ESRC National Centre for Research Methods

BIAS

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Session 52. Combining Data: Quantitative Methods and Applications

Bayesian Graphical Models for Combining Multiple Data Sources

Alexina Mason

Joint work with Sylvia Richardson and Nicky Best

Department of Epidemiology and Biostatistics
Imperial College, London

8 July 2010

http://www.bias-project.org.uk

Outline

Motivation
- Why?
- Case study

Graphical Models
- Introduction to graphical models
- Graphical model for application

Results

Why combine multiple datasets?

- Data for the social and health sciences typically come from observational studies
- Due to the complex nature of the research question, a single data set may not provide sufficient information for valid inference
- Some data sources, such as routinely collected administrative data, have a limited number of variables for a large population
- Others, such as surveys or cohort studies, contain detailed information on a small sample of individuals
- Problems such as bias and small sample size can be mitigated by combining multiple sources of data

Case study:
water disinfection by-products and low birth weight

Objective: to estimate the association between trihalomethane (THM) concentrations, a by-product of chlorine water disinfection potentially harmful for reproductive outcomes, and risk of full term low birthweight (<2.5kg).

- Use information on births between 2000 and 2001 in North West England, serviced by the United Utilities Water Company
- Link birth records to estimated trihalomethane water concentrations using
  - residence at birth
  - a model to estimate THM concentration from the water company monitored samples
- First analysis in Molitor et al (2009)

The primary data: HES

- 8969 birth records were obtained from the Hospital Episode Statistics (HES) data base
- Advantage:
  - captures information on all hospital births in the population under study ⇒ increased power, fully representative
- Disadvantage:
  - contains only limited information on mother and infant characteristics which impact birth weight ⇒ increased bias

The primary data: HES

- 8969 birth records were obtained from the Hospital Episode Statistics (HES) data base
- Advantage:
  - captures information on all hospital births in the population under study ⇒ increased power, fully representative
- Disadvantage:
  - contains only limited information on mother and infant characteristics which impact birth weight ⇒ increased bias
- HES contains data on
  - mother's age
  - baby gender
  - gestational age
  - an index of deprivation
- But no data on other characteristics which impact birth weight
  - maternal smoking
  - ethnicity

A naive analysis using HES data only

- denote low birthweight by Y (binary indicator)
- fit a logistic regression model using the
  - exposure of interest, X=THM
  - measured confounders,
    C={mother's age, baby gender, deprivation index}
- ignore the unmeasured confounders,
  U={maternal smoking, ethnicity}

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$
$$logit(p_i) = \beta_0 + \beta_X X_i + \beta_C^T \boldsymbol{C}_i$$

Analysis results using HES data only (n=8969)

no adjustment for mother’s smoking and ethnicity status

	Odds ratio (95% interval estimate)
Trihalomethanes	
$> 60\mu g/L$	1.39 (1.10,1.76)
Mother’s age	
≤ 25	1.14 (0.86,1.52)
$25-29^{\star}$	1
$30-34$	0.81 (0.57,1.15)
≥ 35	1.10 (0.73,1.65)
Male baby	0.76 (0.60,0.96)
Deprivation index	1.37 (1.20,1.56)

* Reference group

Biased from unmeasured confounders?

The supplementary data: MCS

- The Millennium Cohort Study (MCS)
  - contains survey information on mothers and infants born during 2000-2001
  - includes detailed information on ethnicity and smoking
  - is a stratified sample (advantaged/disadvantage/ethnic minority)
- For our study region, 824 cohort births can be matched to the hospital data (HES)

MCS Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \alpha_{s(i)} + \beta_X X_i + \boldsymbol{\beta}_C^T \mathbf{C}_i + \boldsymbol{\beta}_U^T \mathbf{U}_i$$

The supplementary data: MCS

- The Millennium Cohort Study (MCS)
  - contains survey information on mothers and infants born during 2000-2001
  - includes detailed information on ethnicity and smoking
  - is a stratified sample (advantaged/disadvantage/ethnic minority)
- For our study region, 824 cohort births can be matched to the hospital data (HES)

MCS Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \alpha_{s(i)} + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i + \boldsymbol{\beta}_U^T \boldsymbol{U}_i$$

include stratum specific intercepts, $\alpha_{s(i)}$

The supplementary data: MCS

- The Millennium Cohort Study (MCS)
  - contains survey information on mothers and infants born during 2000-2001
  - includes detailed information on ethnicity and smoking
  - is a stratified sample (advantaged/disadvantage/ethnic minority)
- For our study region, 824 cohort births can be matched to the hospital data (HES)

MCS Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \alpha_{s(i)} + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i + \boldsymbol{\beta}_U^T \boldsymbol{U}_i$$

include stratum specific intercepts, $\alpha_{s(i)}$

add $\boldsymbol{U}$={smoking, ethnicity}

Analysis results using MCS data only (n=824)

	Odds ratio (95% interval estimate)		
	HES only (excludes ***U***)	MCS only (excludes ***U***)	MCS only (includes ***U***)
Trihalomethanes			
$> 60\mu g/L$	1.39 (1.10,1.76)	2.06 (0.85,4.98)	1.87 (0.76, 4.62)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.65 (0.23,1.79)	0.57 (0.20, 1.61)
$25-29^\star$	1	1	1
$30-34$	0.81 (0.57,1.15)	0.13 (0.02,1.11)	0.13 (0.02, 1.11)
≥ 35	1.10 (0.73,1.65)	1.57 (0.49,5.08)	1.82 (0.55, 5.99)
Male baby	0.76 (0.60,0.96)	0.59 (0.25,1.43)	0.62 (0.25, 1.49)
Deprivation index	1.37 (1.20,1.56)	1.54 (0.78,3.02)	1.44 (0.73, 2.85)
Smoking			3.39 (1.26, 9.12)
Non-white ethnicity			2.66 (0.69,10.31)

* Reference group

Lacks power to detect an association

Analysis results using MCS data only (n=824)

	Odds ratio (95% interval estimate)		
	HES only	MCS only	MCS only
	(excludes ***U***)	(excludes ***U***)	(includes ***U***)
Trihalomethanes			
$> 60\mu g/L$	1.39 (1.10,1.76)	2.06 (0.85,4.98)	1.87 (0.76, 4.62)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.65 (0.23,1.79)	0.57 (0.20, 1.61)
$25 - 29^\star$	1	1	1
$30 - 34$	0.81 (0.57,1.15)	0.13 (0.02,1.11)	0.13 (0.02, 1.11)
≥ 35	1.10 (0.73,1.65)	1.57 (0.49,5.08)	1.82 (0.55, 5.99)
Male baby	0.76 (0.60,0.96)	0.59 (0.25,1.43)	0.62 (0.25, 1.49)
Deprivation index	1.37 (1.20,1.56)	1.54 (0.78,3.02)	1.44 (0.73, 2.85)
Smoking			3.39 (1.26, 9.12)
Non-white ethnicity			2.66 (0.69,10.31)

* Reference group

Some evidence of confounding

Combining the HES and MCS data

- The objective is to estimate the association between X and Y while controlling for $(\boldsymbol{C}, \boldsymbol{U})$
- Combining HES and MCS data
  - $\boldsymbol{U}$ becomes a vector of partially measured confounders
  - all the observed values of $\boldsymbol{U}$ come from the MCS
  - converts the missing confounder problem into a missing data problem
- Our modelling strategy is to build a joint model containing
  - an analysis sub-model (to answer question of interest)
  - an imputation sub-model (to impute missing $\boldsymbol{U}$)

Combining the HES and MCS data

- The objective is to estimate the association between X and Y while controlling for $(\boldsymbol{C}, \boldsymbol{U})$
- Combining HES and MCS data
  - $\boldsymbol{U}$ becomes a vector of partially measured confounders
  - all the observed values of $\boldsymbol{U}$ come from the MCS
  - converts the missing confounder problem into a missing data problem
- Our modelling strategy is to build a joint model containing
  - an analysis sub-model (to answer question of interest)
  - an imputation sub-model (to impute missing $\boldsymbol{U}$)

We will now look at how Bayesian graphical models can help with this process

Outline

Motivation

Why?

Case study

Graphical Models

Introduction to graphical models

Graphical model for application

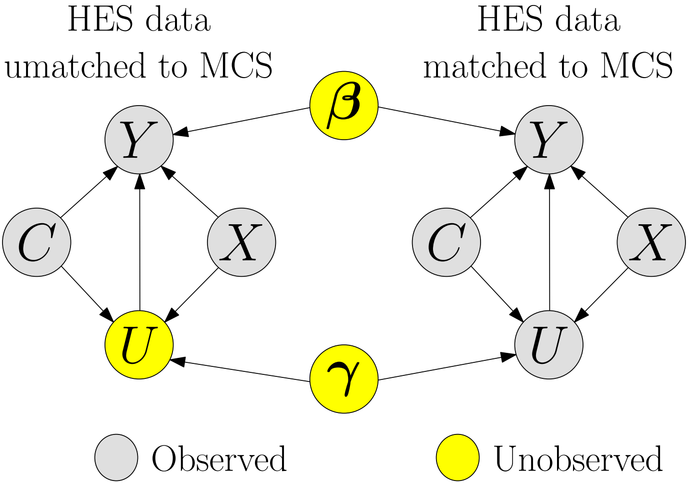
Results

Use of diagrams

Diagrams can be used to visually convey some aspect of a statistical model, for example

Use of diagrams

Diagrams can be used to visually convey some aspect of a statistical model, for example



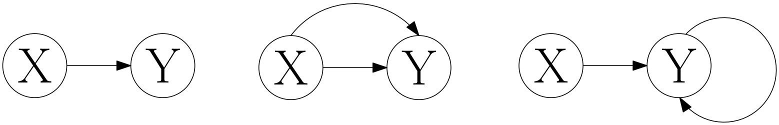
What is a graphical model?

- Diagrams can be used to provide a pictorial representation of
  - the assumed relationships between variables
  - some of the features of the model structure
- Graphical models are formal diagrams that provide a powerful tool for building and communicating complex statistical models
- Formally, a graph, G, consists of
  - finite set of nodes N
  - set of links L, consisting of ordered pairs of distinct elements of N
- G cannot have multiple links or loops

What is a graphical model?

- Diagrams can be used to provide a pictorial representation of
  - the assumed relationships between variables
  - some of the features of the model structure
- Graphical models are formal diagrams that provide a powerful tool for building and communicating complex statistical models
- Formally, a graph, G, consists of
  - finite set of nodes N
  - set of links L, consisting of ordered pairs of distinct elements of N
- G cannot have multiple links or loops

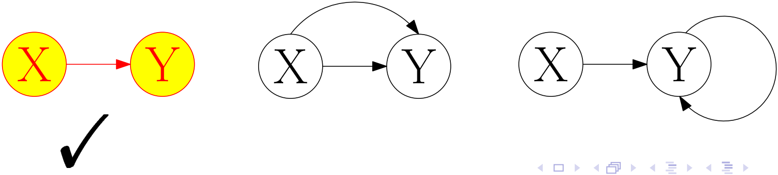
Example: $N = \{X, Y\}$;



What is a graphical model?

- Diagrams can be used to provide a pictorial representation of
  - the assumed relationships between variables
  - some of the features of the model structure
- Graphical models are formal diagrams that provide a powerful tool for building and communicating complex statistical models
- Formally, a graph, G, consists of
  - finite set of nodes N
  - set of links L, consisting of ordered pairs of distinct elements of N
- G cannot have multiple links or loops

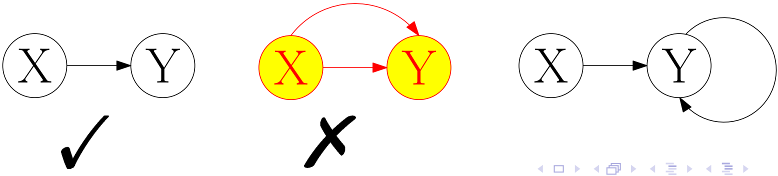
Example: $N = \{X, Y\}$; $L = \{(X, Y)\}$



What is a graphical model?

- Diagrams can be used to provide a pictorial representation of
  - the assumed relationships between variables
  - some of the features of the model structure
- Graphical models are formal diagrams that provide a powerful tool for building and communicating complex statistical models
- Formally, a graph, G, consists of
  - finite set of nodes N
  - set of links L, consisting of ordered pairs of distinct elements of N
- G cannot have multiple links or loops

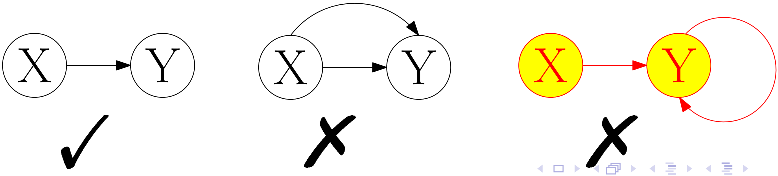
Example: $N = \{X, Y\}$; $L = \{(X, Y), (X, Y)\}$



What is a graphical model?

- Diagrams can be used to provide a pictorial representation of
  - the assumed relationships between variables
  - some of the features of the model structure
- Graphical models are formal diagrams that provide a powerful tool for building and communicating complex statistical models
- Formally, a graph, G, consists of
  - finite set of nodes N
  - set of links L, consisting of ordered pairs of distinct elements of N
- G cannot have multiple links or loops

Example: $N = \{X, Y\}$; $L = \{(X, Y), (Y, Y)\}$



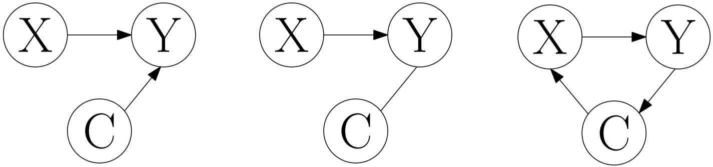
Introducing the Directed Acyclic Graph

- A Directed Acyclic Graph (DAG) is a particular type of graphical model which contains
  - only directed links
  - no cycles

Introducing the Directed Acyclic Graph

- A Directed Acyclic Graph (DAG) is a particular type of graphical model which contains
  - only directed links
  - no cycles

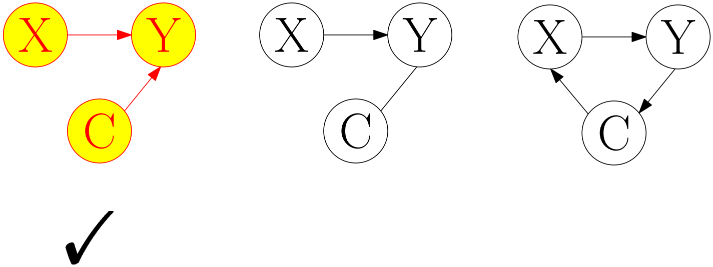
Example: $N = \{X, Y, C\}$;



Introducing the Directed Acyclic Graph

- A Directed Acyclic Graph (DAG) is a particular type of graphical model which contains
  - only directed links
  - no cycles

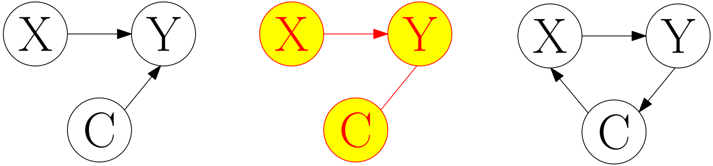
Example: $N = \{X, Y, C\}$; $L = \{(X, Y), (C, Y)\}$



Introducing the Directed Acyclic Graph

- A Directed Acyclic Graph (DAG) is a particular type of graphical model which contains
  - only directed links
  - no cycles

Example: $N = \{X, Y, C\}$; $L = \{(X, Y), (C, Y), (Y, C)\}$



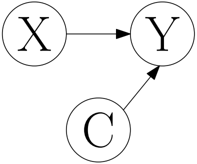
✓ ✗

undirected link

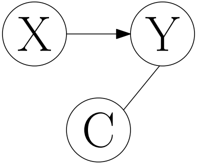
Introducing the Directed Acyclic Graph

- A Directed Acyclic Graph (DAG) is a particular type of graphical model which contains
  - only directed links
  - no cycles

Example: $N = \{X, Y, C\}$; $L = \{(X, Y), (Y, C), (C, X)\}$

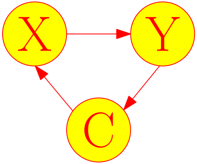


✓



✗

undirected link



✗

cycle

How to represent a statistical model using a DAG

- Each element of the statistical model (variables, parameters, etc.) is represented by a node
- The assumed relationships between these elements are represented by links
- The direction of the links reflects the dependence implied by the equations
- However, the arrows have no intrinsic meaning - they should NOT be interpreted as meaning causal inference unless extra assumptions are made

How to represent a statistical model using a DAG

- Each element of the statistical model (variables, parameters, etc.) is represented by a node
- The assumed relationships between these elements are represented by links
- The direction of the links reflects the dependence implied by the equations
- However, the arrows have no intrinsic meaning - they should NOT be interpreted as meaning causal inference unless extra assumptions are made

We now provide examples using our application

DAG for analysis sub-model

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \beta_C^T \boldsymbol{C}_i$$

$nodes = \{\beta, X_i, \boldsymbol{C}_i, Y_i, p_i\}$

$links = \{(p_i, Y_i), (\beta, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i)\}$

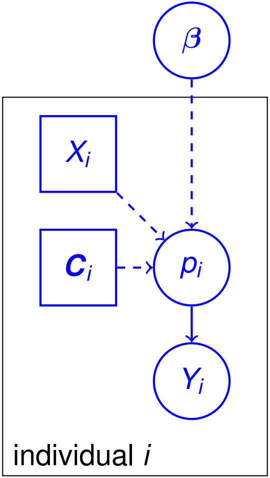
circular nodes denote random variables

square nodes denote constant quantities

probabilistic links indicated by solid arrows ("$\sim$")

deterministic links indicated by dashed arrows ("=")

repeated structure indicated by a rectangle, "plate"



DAG for analysis sub-model

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i$$

$nodes = \{\boldsymbol{\beta}, X_i, \boldsymbol{C}_i, Y_i, p_i\}$

$links = \{(p_i, Y_i), (\boldsymbol{\beta}, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i)\}$

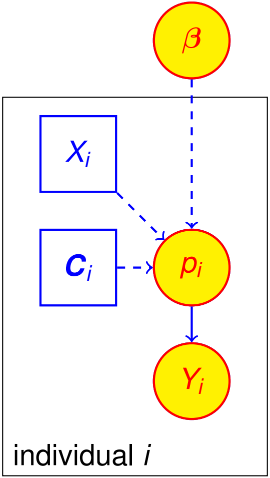
circular nodes denote random variables

square nodes denote constant quantities

probabilistic links indicated by solid arrows ("$\sim$")

deterministic links indicated by dashed arrows ("=")

repeated structure indicated by a rectangle, "plate"



DAG for analysis sub-model

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \beta_C^T \boldsymbol{C}_i$$

$nodes = \{\beta, X_i, \boldsymbol{C}_i, Y_i, p_i\}$

$links = \{(p_i, Y_i), (\beta, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i)\}$

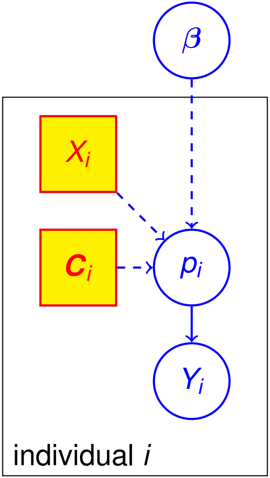
circular nodes denote random variables

square nodes denote constant quantities

probabilistic links indicated by solid arrows ("$\sim$")

deterministic links indicated by dashed arrows ("=")

repeated structure indicated by a rectangle, "plate"



DAG for analysis sub-model

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \beta_C^T \boldsymbol{C}_i$$

$nodes = \{\beta, X_i, \boldsymbol{C}_i, Y_i, p_i\}$

$links = \{(p_i, Y_i), (\beta, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i)\}$

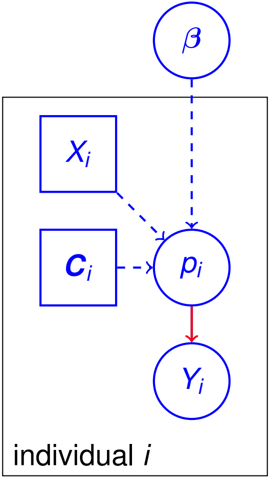
circular nodes denote random variables

square nodes denote constant quantities

probabilistic links indicated by solid arrows ("$\sim$")

deterministic links indicated by dashed arrows ("=")

repeated structure indicated by a rectangle, "plate"



DAG for analysis sub-model

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i$$

$nodes = \{\boldsymbol{\beta}, X_i, \boldsymbol{C}_i, Y_i, p_i\}$

$links = \{(p_i, Y_i), (\boldsymbol{\beta}, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i)\}$

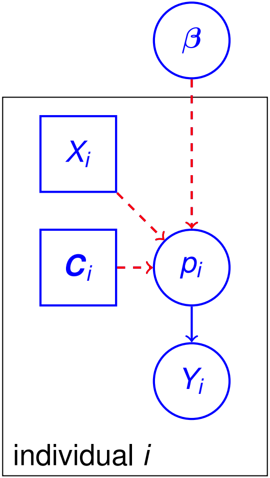
circular nodes denote random variables

square nodes denote constant quantities

probabilistic links indicated by solid arrows ("$\sim$")

deterministic links indicated by dashed arrows ("=")

repeated structure indicated by a rectangle, "plate"



DAG for analysis sub-model

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i$$

$nodes = \{\boldsymbol{\beta}, X_i, \boldsymbol{C}_i, Y_i, p_i\}$

$links = \{(p_i, Y_i), (\boldsymbol{\beta}, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i)\}$

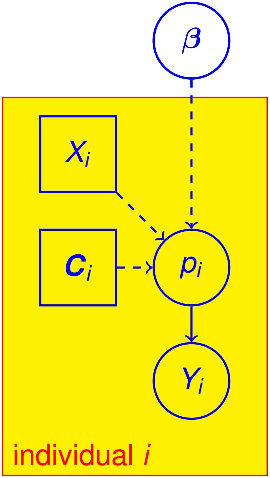
circular nodes denote random variables

square nodes denote constant quantities

probabilistic links indicated by solid arrows ("$\sim$")

deterministic links indicated by dashed arrows ("=")

repeated structure indicated by a rectangle, "plate"



DAG for analysis sub-model

Naive Analysis Model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i$$

$$nodes = \{\boldsymbol{\beta}, X_i, \boldsymbol{C}_i, Y_i, p_i\}$$

$$links = \{(p_i, Y_i), (\boldsymbol{\beta}, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i)\}$$

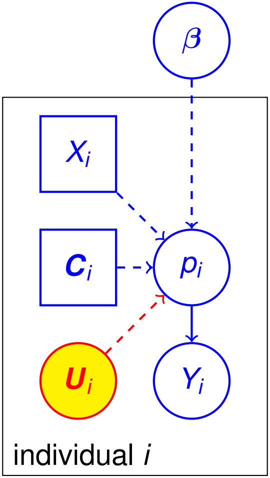
Analysis sub-model

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i + \boldsymbol{\beta}_U^T \boldsymbol{U}_i$$

$$N = \{\boldsymbol{\beta}, X_i, \boldsymbol{C}_i, \boldsymbol{U}_i, Y_i, p_i\}$$

$$L = \{(p_i, Y_i), (\boldsymbol{\beta}, p_i), (X_i, p_i), (\boldsymbol{C}_i, p_i), (\boldsymbol{U}_i, p_i)\}$$



Accounting for sampling bias

Now turning to the imputation sub-model

- The supplementary data (MCS) is not a random sample from the primary data (HES)
- The MCS cohort sampling was stratified in order to oversample low socio-economic and ethnic categories
- To account for the sampling bias, include the stratum in the imputation model as stratum specific intercepts, $\alpha_{s(i)}$

Imputation sub-model

- Missing values of U (smoking and ethnicity) imputed using latent variables, $U^{\star}$
- Allows for correlation between U
- Multivariate probit model is defined as follows:

$$\boldsymbol{U}_i^{\star} \sim MVN(\boldsymbol{\mu}_i, \Sigma)$$

$$\boldsymbol{\mu}_i = \alpha_{s(i)} + \gamma_X X_i + \gamma_C^T \boldsymbol{C_i}$$

$$U_{ij} = I(U_{ij}^{\star} > 0),\ j = 1,2$$

$$\boldsymbol{U}_i^{\star} = \begin{pmatrix} U_{i1}^{\star} \\ U_{i2}^{\star} \end{pmatrix},\ \boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix},\ \Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

DAG for imputation sub-model

Imputation sub-model

$$\boldsymbol{U}_i^\star \sim MVN(\boldsymbol{\mu}_i, \Sigma)$$

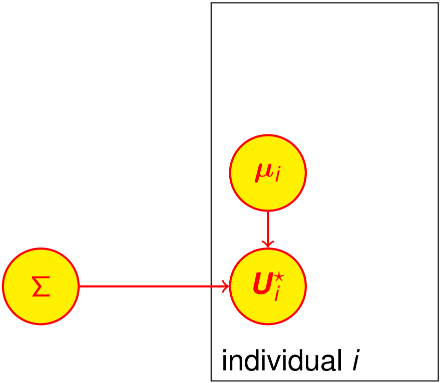
$$\boldsymbol{\mu}_i = \alpha_{s(i)} + \gamma_X X_i + \gamma_C^T \boldsymbol{C}_i$$

$$U_{ij} = I(U_{ij}^\star > 0)\ j = 1, 2$$

$$\boldsymbol{U}_i^\star = \begin{pmatrix} U_{i1}^\star \\ U_{i2}^\star \end{pmatrix}$$

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$



DAG for imputation sub-model

Imputation sub-model

$$\boldsymbol{U}_i^\star \sim MVN(\boldsymbol{\mu}_i, \Sigma)$$

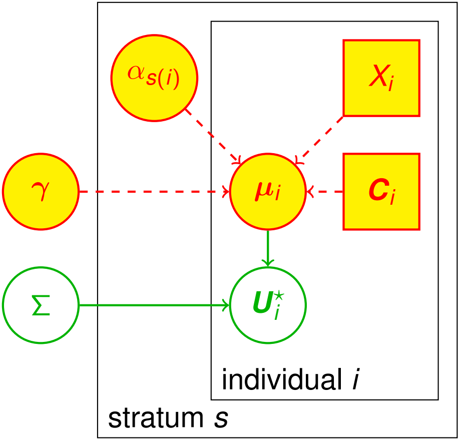
$$\boldsymbol{\mu}_i = \alpha_{s(i)} + \gamma_X X_i + \gamma_C^T \boldsymbol{C_i}$$

$$U_{ij} = I(U_{ij}^\star > 0)\ j = 1,2$$

$$\boldsymbol{U}_i^\star = \begin{pmatrix} U_{i1}^\star \\ U_{i2}^\star \end{pmatrix}$$

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$



DAG for imputation sub-model

Imputation sub-model

$\boldsymbol{U}_i^\star \sim MVN(\boldsymbol{\mu}_i, \Sigma)$

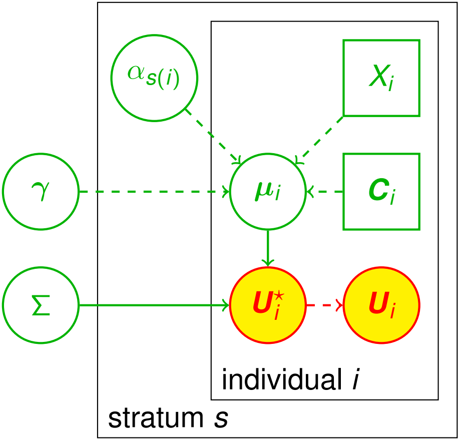
$\boldsymbol{\mu}_i = \alpha_{s(i)} + \gamma_X X_i + \gamma_C^T \boldsymbol{C}_i$

$U_{ij} = I(U_{ij}^\star > 0)\ j = 1, 2$

$\boldsymbol{U}_i^\star = \begin{pmatrix} U_{i1}^\star \\ U_{i2}^\star \end{pmatrix}$

$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}$

$\Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$



DAG for imputation sub-model

Imputation sub-model

$$\boldsymbol{U}_i^\star \sim MVN(\boldsymbol{\mu}_i, \Sigma)$$

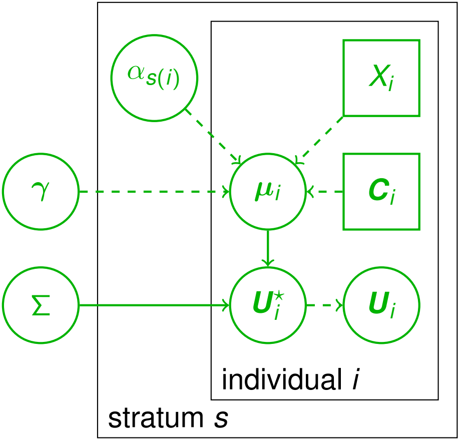
$$\boldsymbol{\mu}_i = \alpha_{s(i)} + \gamma_X X_i + \gamma_C^T \boldsymbol{C_i}$$

$$U_{ij} = I(U_{ij}^\star > 0)\ j = 1, 2$$

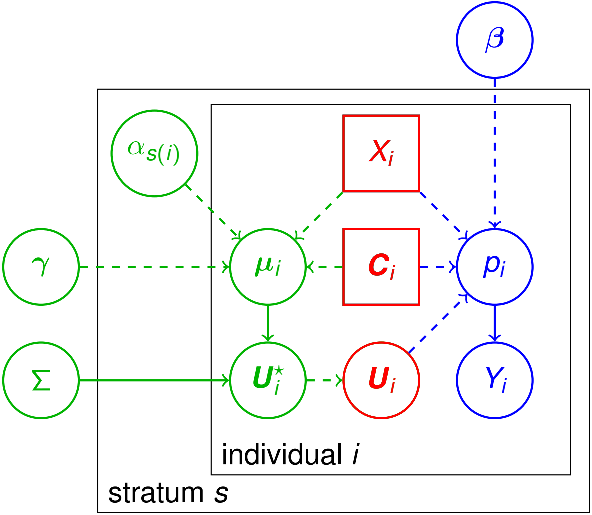
$$\boldsymbol{U}_i^\star = \begin{pmatrix} U_{i1}^\star \\ U_{i2}^\star \end{pmatrix}$$

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}$$

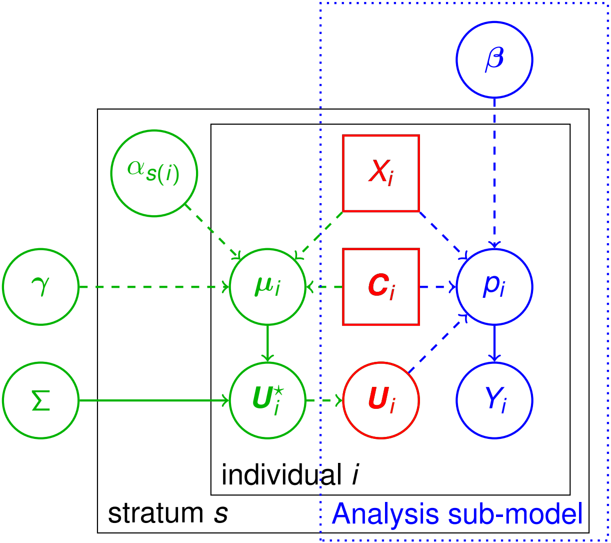
$$\Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$



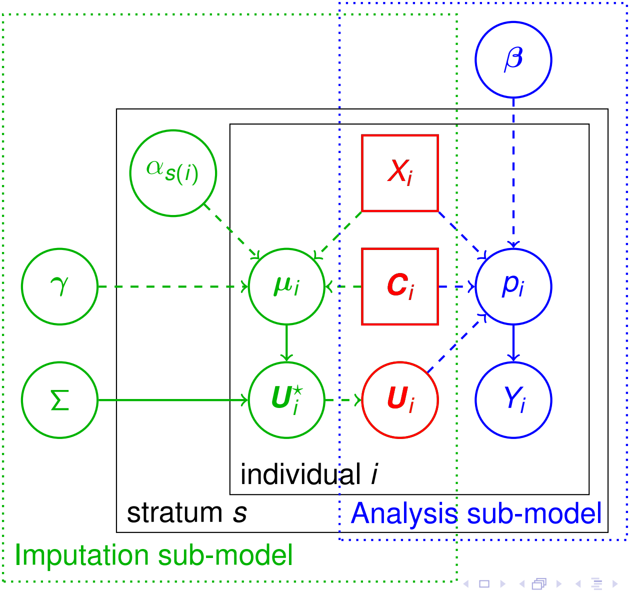
DAG for joint model



DAG for joint model



DAG for joint model



The power of graphical models

- DAGs show how sub-models fit together to form an overall model
- But, there is far more to a DAG than visual representation
- DAGs encode conditional independence statements, which
  - allow a joint distribution to be decomposed into a product of conditional distributions (factorisation theorem)
  - is very useful for implementing Markov chain Monte Carlo (MCMC) methods for Bayesian inference (e.g. exploited by WinBUGS software)
  - ensures joint model is consistent

Comparison with conventional Multiple Imputation by Chained Equations (MICE)

- Specify a separate conditional distribution for each variable with missing data: U_1 = smoking; U_2 = ethnicity
- No longer uses latent variables

MICE imputation model

$$
\begin{aligned}
U_{1i} &\sim Bernoulli(q_i) \\
logit(q_i) &= \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i \\
U_{2i} &\sim Bernoulli(r_i) \\
logit(r_i) &= \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i
\end{aligned}
$$

Comparison with conventional Multiple Imputation by Chained Equations (MICE)

- Specify a separate conditional distribution for each variable with missing data: U_1 = smoking; U_2 = ethnicity
- No longer uses latent variables

MICE imputation model

$$U_{1i} \sim Bernoulli(q_i)$$
$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$
$$U_{2i} \sim Bernoulli(r_i)$$
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

include other U

Comparison with conventional Multiple Imputation by Chained Equations (MICE)

- Specify a separate conditional distribution for each variable with missing data: U_1 = smoking; U_2 = ethnicity
- No longer uses latent variables

MICE imputation model

$$
\begin{aligned}
U_{1i} &\sim Bernoulli(q_i) \\
logit(q_i) &= \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i \\
U_{2i} &\sim Bernoulli(r_i) \\
logit(r_i) &= \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i
\end{aligned}
$$

include other U

include response

Comparison with conventional Multiple Imputation by Chained Equations (MICE)

- Specify a separate conditional distribution for each variable with missing data: U_1 = smoking; U_2 = ethnicity
- No longer uses latent variables

MICE imputation model

$$
\begin{aligned}
U_{1i} &\sim Bernoulli(q_i) \\
logit(q_i) &= \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i \\
U_{2i} &\sim Bernoulli(r_i) \\
logit(r_i) &= \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i
\end{aligned}
$$

include other *U*

include response

How can we represent this model graphically?

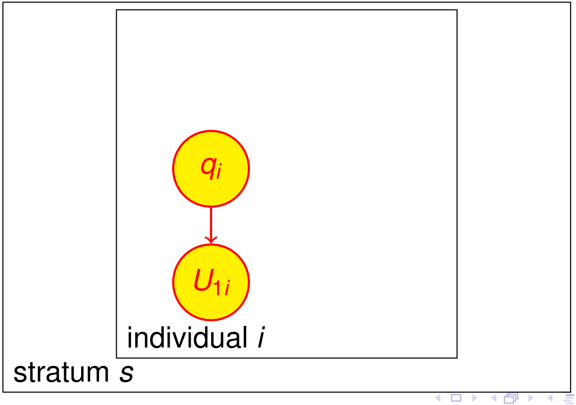
Graphical representation for MICE approach

$U_{1i} \sim Bernoulli(q_i)$

$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$

$U_{2i} \sim Bernoulli(r_i)$

$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$



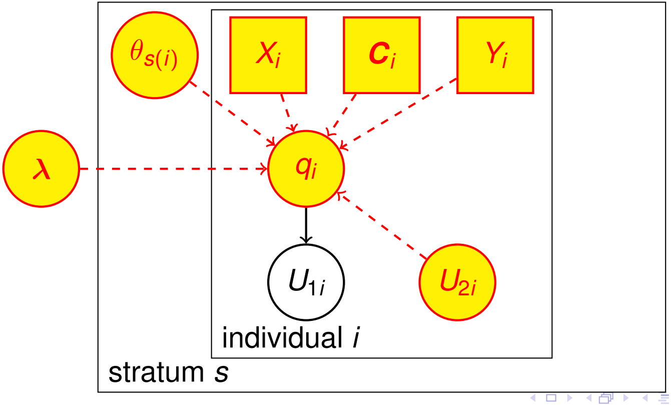
Graphical representation for MICE approach

$U_{1i} \sim Bernoulli(q_i)$

$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$

$U_{2i} \sim Bernoulli(r_i)$

$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$



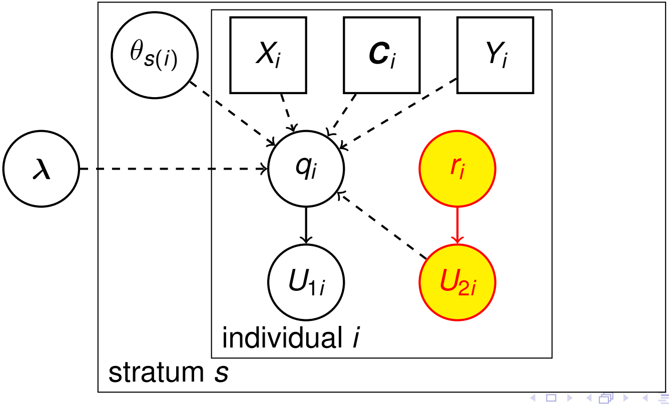
Graphical representation for MICE approach

$U_{1i} \sim Bernoulli(q_i)$

$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$

$U_{2i} \sim Bernoulli(r_i)$

$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$



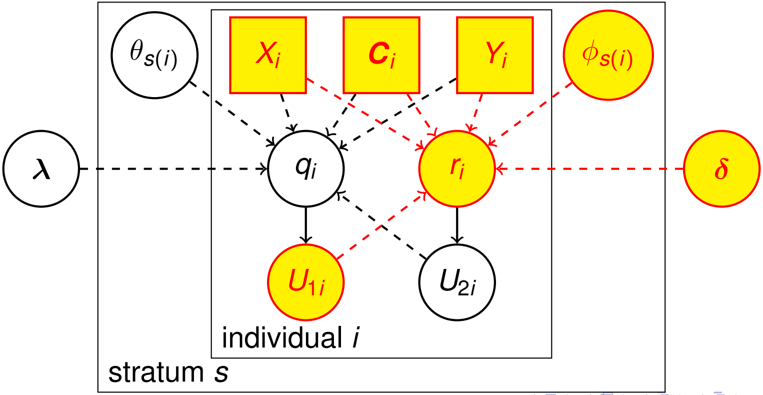
Graphical representation for MICE approach

$U_{1i} \sim Bernoulli(q_i)$

$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$

$U_{2i} \sim Bernoulli(r_i)$

$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$



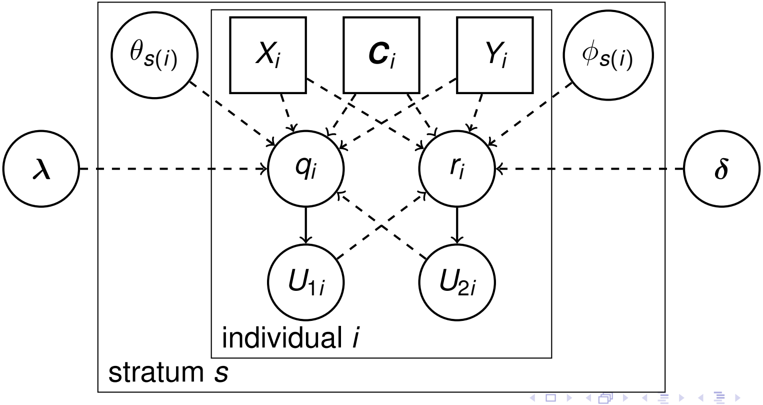
Graphical representation for MICE approach

$$U_{1i} \sim Bernoulli(q_i)$$

$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$

$$U_{2i} \sim Bernoulli(r_i)$$

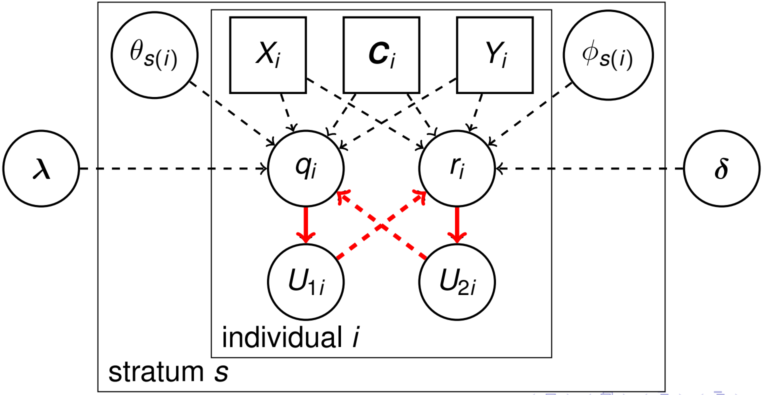
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$



Graphical representation for MICE approach

$$U_{1i} \sim Bernoulli(q_i)$$
$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$
$$U_{2i} \sim Bernoulli(r_i)$$
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

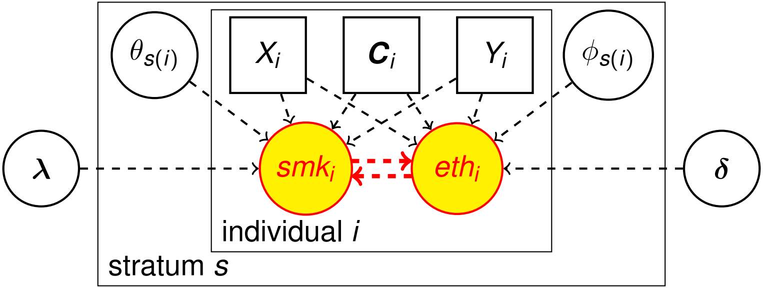
We have a cycle so diagram is NOT a DAG!



Graphical representation for MICE approach

$$U_{1i} \sim Bernoulli(q_i)$$
$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$
$$U_{2i} \sim Bernoulli(r_i)$$
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

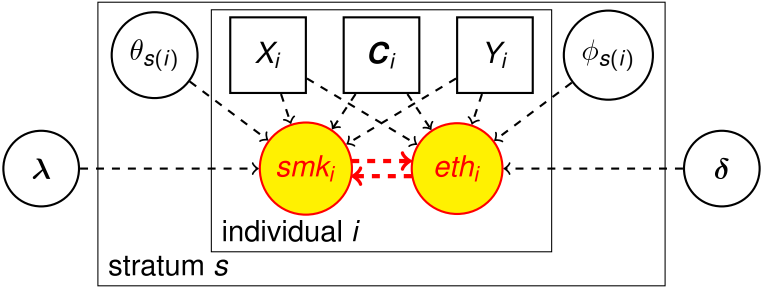
We have a cycle so diagram is NOT a DAG!



Graphical representation for MICE approach

$$U_{1i} \sim Bernoulli(q_i)$$
$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \mathbf{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$
$$U_{2i} \sim Bernoulli(r_i)$$
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \mathbf{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

We have a cycle so diagram is NOT a DAG!



It is not even a graphical model

We iterate between 2 parts of imputation model, then fit analysis model

Graphical representation for MICE approach

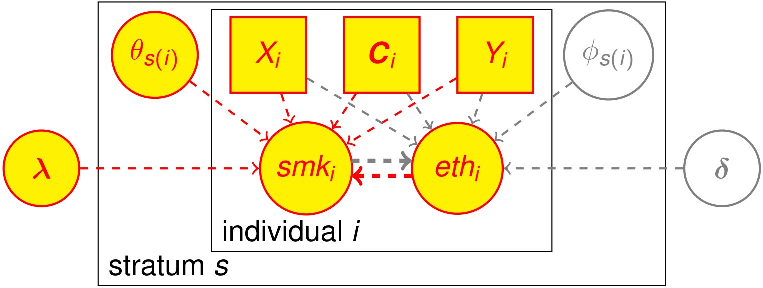
$$U_{1i} \sim Bernoulli(q_i)$$

$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$

$$U_{2i} \sim Bernoulli(r_i)$$

$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

We have a cycle so diagram is NOT a DAG!



It is not even a graphical model

We iterate between 2 parts of imputation model, then fit analysis model

Graphical representation for MICE approach

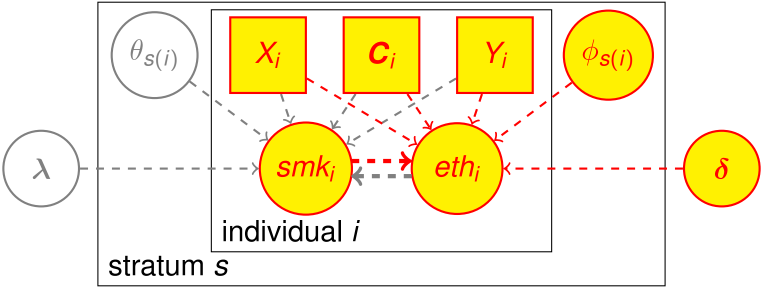
$$U_{1i} \sim Bernoulli(q_i)$$

$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$

$$U_{2i} \sim Bernoulli(r_i)$$

$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

We have a cycle so diagram is NOT a DAG!



It is not even a graphical model

We iterate between 2 parts of imputation model, then fit analysis model

Outline

Motivation

- Why?
- Case study

Graphical Models

- Introduction to graphical models
- Graphical model for application

Results

Comparison of results

	Odds ratio (95% interval estimate)		
	HES only	MCS only	HES+MCS
Trihalomethanes			
$> 60\mu g/L$	1.39 (1.10,1.76)	1.87 (0.76, 4.62)	1.17 (0.88,1.53)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.57 (0.20, 1.61)	1.02 (0.71,1.38)
$25 - 29^\star$	1	1	1
$30 - 34$	0.81 (0.57,1.15)	0.13 (0.02, 1.11)	0.85 (0.57,1.21)
≥ 35	1.10 (0.73,1.65)	1.82 (0.55, 5.99)	1.43 (0.88,2.21)
Male baby	0.76 (0.60,0.96)	0.62 (0.25, 1.49)	0.76 (0.59,0.97)
Deprivation index	1.37 (1.20,1.56)	1.44 (0.73, 2.85)	1.19 (1.01,1.38)
Smoking		3.39 (1.26, 9.12)	3.91 (1.35,9.92)
Non-white ethnicity		2.66 (0.69,10.31)	3.56 (1.75,6.82)

* Reference group

Accounting for missing confounders has reduced OR of THM

Comparison of results II

	Odds ratio (95% interval estimate)		
	HES only	HES+MCS (joint model)	HES+MCS (MICE: 5 imputations)
Trihalomethanes			
$> 60 \mu g/L$	1.39 (1.10,1.76)	1.17 (0.88,1.53)	1.22 (0.91, 1.62)
Mother's age			
≤ 25	1.14 (0.86,1.52)	1.02 (0.71,1.38)	0.98 (0.69, 1.38)
$25 - 29^{\star}$	1	1	1
$30 - 34$	0.81 (0.57,1.15)	0.85 (0.57,1.21)	0.84 (0.58, 1.22)
≥ 35	1.10 (0.73,1.65)	1.43 (0.88,2.21)	1.32 (0.86, 2.03)
Male baby	0.76 (0.60,0.96)	0.76 (0.59,0.97)	0.73 (0.58, 0.93)
Deprivation index	1.37 (1.20,1.56)	1.19 (1.01,1.38)	1.23 (1.05, 1.44)
Smoking		3.91 (1.35,9.92)	4.01 (1.32,12.15)
Non-white ethnicity		3.56 (1.75,6.82)	2.73 (1.83, 4.09)

* Reference group

MICE also reduces OR of THM, but greater uncertainty

Concluding remarks

- Bayesian graphical models are a powerful and flexible tool for building realistic models for complex problems
- Bayesian graphical models
    - allow complex models to be built from smaller comprehensible pieces
    - allow formal combining of multiple data sources
    - result in principled inference
    - ensure all sources of uncertainty are automatically propagated

Further Information

- Coming soon:
  - paper on comparisons of different imputation strategies
    see BIAS web site (www.bias-project.org.uk)
  - introduction to graphical models training materials
    see LEMMA multilevel modelling on-line learning course (www.cmm.bristol.ac.uk/research/Lemma)

▶ Molitor, N.-T., Best, N., Jackson, C., and Richardson, S. (2009).
Using Bayesian graphical models to model biases in observational studies and to combine multiple data sources: Application to low birth-weight and water disinfection by-products.
Journal of the Royal Statistical Society, Series A (Statistics in Society), **172**, (3), 615–37.

Acknowledgements

- Funding by ESRC
- The BIAS project (PI N Best), based at Imperial College, London, is a node of the Economic and Social Research Council's National Centre for Research Methods (NCRM)
- For papers and technical reports, see our web site www.bias-project.org.uk