Multilevel Analysis

Further Statistical Methods, Lecture 8
Hilary Term 2006

Steffen Lauritzen, University of Oxford; February 15, 2006

Multilevel observations

Multilevel analysis is concerned with observations with a *nested* structure.

For a two-level analysis we typically think of *individuals* within *groups*. The individual level is in general called *level one*, the group level *level two*.

An example of observations of this type can for example be performance measures for *pupils* of a specific age-group within *classes*.

The levels could be nested yet another time as e.g. classes within *schools*. And further, the schools could be grouped according to *regions* within *countries*, etc. although at the

top-level there might well be problems of compatibility of performance measures.

For simplicity we will only consider two levels, pupils within classes.

An example

As our basic example we will consider a Dutch study comprising $N = 131$ classes, each of sizes between 4 and 35, with a total of $M = 2287$ pupils.

The performance measure of interest is the score on a language test, and explanatory variables include class sizes and the IQ of individual pupils.

We let $Y_{ij}, j = 1, \ldots N, i = 1, \ldots n_j$ be the score for pupil i in class j and study the dependence of this response on covariates such as the IQ x_{ij} of the pupil and the size z_j of the class.

x_{ij} are *level one covariates* and z_j *level two covariates*.

A simple regression model

A first attempt could be to let

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + R_{ij}$$

with R_{ij} independent and distributed as $\mathcal{N}(0, \sigma^2)$.

This is a standard linear regression model which only has an indirect multilevel character.

The model ignores that pupils in the same class will tend to have more similar scores than those in different classes, even when the covariates are taken into account.

This is a *very serious mistake* if the variations in score at group level are not fully explained by the covariates.

Introducing random effects

For a moment, ignore the covariates x_{ij} and z_j and consider instead the model

$$Y_{ij} = \beta_0 + U_j + R_{ij}$$

where $U_j \sim \mathcal{N}(0, \tau^2)$. This model then has

$$\mathbf{V}(Y_{ij}) = \sigma^2 + \tau^2, \quad \mathrm{Cov}(Y_{ij}, Y_{i'j}) = \tau^2, \quad \mathrm{Cov}(Y_{ij}, Y_{i'j'}) = 0$$

so that scores of pupils within the same class are correlated. The correlation is

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

and is known as the *intraclass correlation coefficient.*

This type of model is also known as a *random effects model* since one could think of $\beta_j = \beta_0 + U_j$ as a group effect, in this case modelled as a random effect. Adding back the covariates leads to

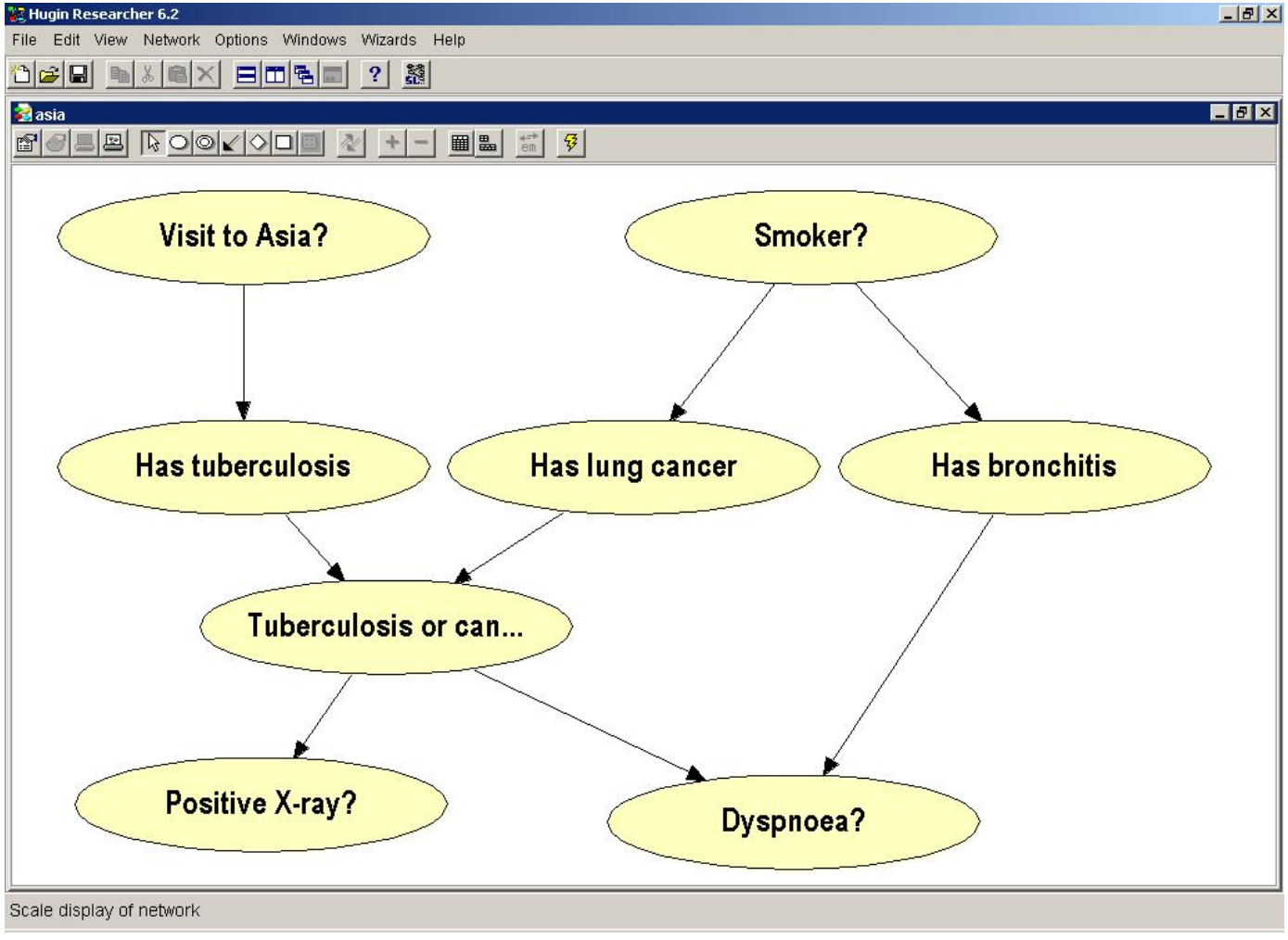
$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + U_j + R_{ij}.$$

It can give a better overview to introduce an intermediate variable describing the total class effect

$$M_j = \beta_0 + \beta_2 z_j + U_j; \quad Y_{ij} = M_j + \beta_1 x_{ij} + R_{ij}$$

where M_j now become missing data, or rather latent variables.

Example of a directed graphical model



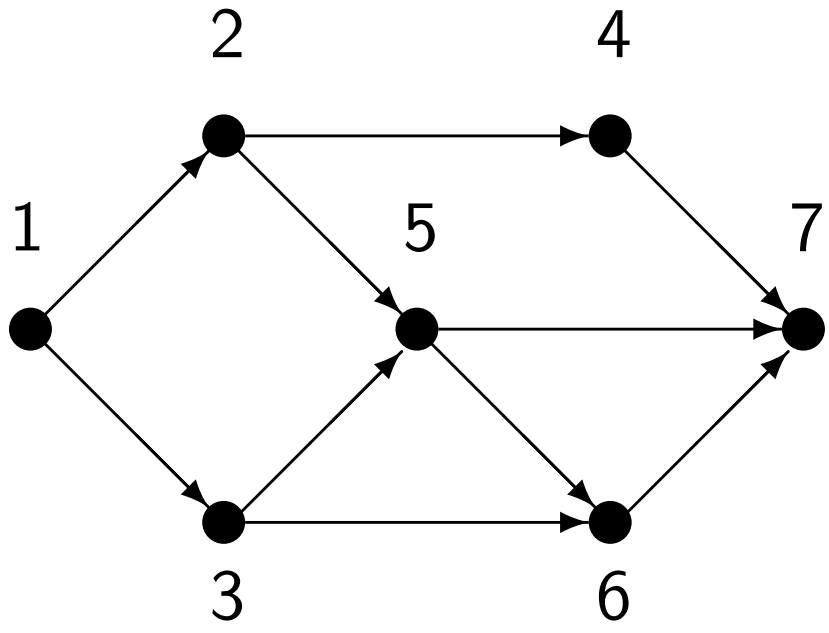
Directed graphical models

A probability distribution *factorizes* w.r.t. a *directed acyclic graph* (DAG) $\mathcal{D}$ if it has density or probability mass function f of the form

$$f(x) = \prod_{v \in V} f(x_v \,|\, x_{\mathrm{pa}(v)}),$$

i.e. into a product of the conditional distributions of each node given its parents.

Example of DAG factorization



The above graph corresponds to the factorization

$$\begin{aligned} f(x) &= f(x_1)f(x_2\,|\,x_1)f(x_3\,|\,x_1)f(x_4\,|\,x_2) \\ &\times \quad f(x_5\,|\,x_2,x_3)f(x_6\,|\,x_3,x_5)f(x_7\,|\,x_4,x_5,x_6). \end{aligned}$$

Including parameters in the graph

Directed graphical models become particularly useful when *parameters are explicitly included* in the graph.

The factorization can then be written as

$$f(x \mid \theta) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v),\theta}).$$

Each conditional distribution may only depend of part of the parameter, the 'parameter parents'.

Including observations in the graph

To be able to describe complex observational patterns, we would wish to represent repeated structures. This can be done through *plates* as in WinBUGS.

Estimation of parameters

The maximum likelihood (ML) estimates of the parameters can be obtained using the EM algorithm, treating M_j as missing variables.

For 'complete data', with M_j observed, the estimation problem splits into two simple linear regression problems

1. Estimating $(\beta_0, \beta_2, \tau^2)$ by regressing M_j on z_j;
2. Estimating β_1, σ^2 by regressing $Y_{ij} - M_j$ on x_{ij}

Unfortunately the ML estimates of the variance components (σ^2, τ^2) can be very biased, as these do not

take into account the loss in degrees of freedom due to the estimation of regression coefficients.

Instead a method known as *residual maximum likelihood* or REML is often used.

This involves (in principle) the following steps

1. Calculate initial estimates of regression coefficients using OLS, ignoring the multi-level structure;
2. Form residuals

$$\hat{r}_{ij} = y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_{ij} - \hat{\beta}_2 z_j.$$

3. These residuals $\hat{R}$ follow a multivariate normal distribution with mean 0 and a covariance matrix $\Sigma(\sigma^2, \tau^2)$;

4. The REML estimates of (σ^2, τ^2) are the maximum likelihood estimates *based on the residuals* $\hat{R}$.

5. Revised estimates of the regression parameters are then calculated using appropriate weighted least squares.

An algorithm of EM type exists for calculating the REML estimates, but this and other methods have also been implemented in generally available software.

Estimating random effects

It could be of independent interest, for example when making performance ranking, to estimate the level two effects which are not explained by covariates, i.e.

$$\beta_j = \beta_0 + U_j.$$

This can be done by calculating

$$\hat{\beta}_j = \hat{\beta}_0 + \hat{\mathbf{E}}(U_j \,|\, Y),$$

i.e. the estimated conditional expectation given the observed data.

A Bayesian alternative

An alternative method of analysis is to specify prior distributions of the unknown parameters.

The resulting model is then a *Bayesian hierarchical model.*

It has a simple representation as a Bayesian graphical model and WinBUGS provides the necessary software for estimating all relevant effects using Markov chain Monte-Carlo methods (MCMC).

Beware that prior distributions can be influential.

Note in particular that the parameters mean different things when covariates are centered in different ways, yielding different models with default prior specifications:

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad \alpha \sim N(0, 100), \quad \beta \sim N(0, 100)$$

is very different from

$$Y_i \sim N(\alpha + \beta x_i^*, \sigma^2), \quad \alpha \sim N(0, 100), \quad \beta \sim N(0, 100),$$

where $x_i^* = x_i - \bar{x}$. Without the prior specifications, the models would be equivalent, only the interpretation of α would be different.

Snijders and Bosker (1999) write that BUGS needs balanced data, i.e. equal group sizes, to be applied.

This is not correct, on the contrary, BUGS was developed to allow very unbalanced designs indeed.