Fields Institute, Toronto: September 2004

D.R.Cox

Nuffield College and Department of Statistics, Oxford

david.cox@nuf.ox.ac.uk

Joint work with

Nanny Wermuth

Chalmers/Gothenburg University, Sweden

wermuth@math.chalmers.se

andom variables have finite variance and without loss of

ro mean. Then we can always write

$$Y = \beta X + \epsilon_{Y.X}$$

$\epsilon_{Y.X}) = 0$, called the linear least squares regression

Y on X. Of course its statistical usefulness may be

e defining condition

$$\beta = \mathrm{cov}(Y, X)\{\mathrm{cov}(X, X)\}^{-1}.$$

res property is easily verified.

More on linear least squares regression

More generally if Y and X are vectors we can regress each component of Y on X and require the error to be uncorrelated with all the components of X to obtain

$$Y = \mathrm{B}X + \epsilon,$$

Where, with $\Sigma_{YX} = E(YX^T), \Sigma_{XX} = E(XX^T)$

$$\mathrm{B} = \Sigma_{YX}\Sigma_{XX}^{-1}.$$

Concentration matrices

Write $W = \Sigma_{YY}^{-1} Y$ so that

$$\mathrm{cov}(W, W) = \Sigma_{YY}^{-1}, \mathrm{cov}(Y, W) = I.$$

Thus in the equation

$$W_1 = \sigma^{11} Y_1 + \sigma^{12} Y_2 + \ldots + \sigma^{1d} Y_d$$

W_1 is uncorrelated with every Y_j except Y_1. That is,

$$Y_1 = (-\sigma^{12}/\sigma^{11}) Y_2 + \ldots + (-\sigma^{1d}/\sigma^{11}) Y_d + W_1/\sigma^{11}$$

is a linear least squares regression equation. Thus

$$\rho_{ij.V\setminus i,j} = -\sigma^{ij}/(\sigma^{ii}\sigma^{jj})^{1/2}.$$

Partial and total regression coefficients

Use notation of Yule that shows in a regression coefficient what other variables are involved, i.e. linearly conditioned on. Thus with three variables Y, X, U we write

$$Y = \beta_{YX.U}X + \beta_{YU.X}U + \epsilon_{Y.XU},$$
$$U = \beta_{UX}X + \epsilon_{U.X}.$$

Then directly (Cochran, 1938)

$$\beta_{YX} = \beta_{YX.U} + \beta_{YU.X}\beta_{UX}$$

Gradient analogue

If $y = y(x, u)$ then

$$Dy/Dx = \partial y/\partial x + (\partial y/\partial u)\,(du/dx).$$

Compare with

$$\beta_{YX} = \beta_{YX.U} + \beta_{YU.X}\beta_{UX}$$

The generality of the gradient result suggests that the probabilistic version can be extended.

Also direct extensions to vector Y, X, U.

A fairly general formulation

$$F_{Y|X}(y;x) = \int F_{Y|XU}(y;x,u)d_u F_{U|X}(u;x).$$

Suppose X continuous. Then simplifying the notation slightly

$$\partial F_{Y|X}/\partial x = \int (\partial F_{Y|XU)}/\partial x d_u F_{U|X} + F_{Y|XU}\partial d_u F_{U|X}/\partial x).$$

Integrate the second term by parts and assume regular behaviour at the terminals to give

$$\begin{aligned}\partial F_{Y|X}/\partial x = \int (&\partial F_{Y|XU}/\partial x d_u F_{U|X}\\ &-\partial F_{Y|XU}/\partial u\, \partial F_{U|X}/\partial x)du.\end{aligned}$$

Quantile regression

Define the ϵ point of the conditional distribution of Y given X by

$$F_{Y|X}(y^{\epsilon}(x); x) = \epsilon.$$

Differentiate with respect to x at fixed ϵ. Then

$$F_{Y|X}(y^{\epsilon}(x); x) dy^{\epsilon}(x)/dx + \partial F_{Y|X}(y^{\epsilon}(x); x)/\partial x = 0.$$

Define

$$\gamma_{YX}(y; x) = -\frac{1}{f_{Y|X}(y; x)} \frac{\partial F_{Y|X}(y; x)}{\partial x},$$

etc.

Quantile regression ctd

Thus

$$\gamma_{YX}(y;x) = \int \{\gamma_{YX.U}(y;x,u) + \gamma_{YU.X}(y;u,x)\gamma_{UX}(u;x)\} \\ f_{U|YX}(u;y,x)du.$$

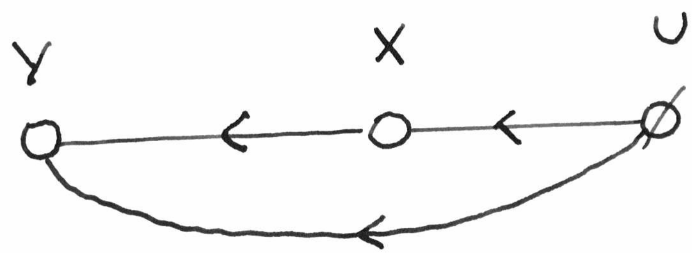
Compare with

$$\beta_{YX} = \beta_{YX.U} + \beta_{YU.X}\beta_{UX}.$$

Another implication of formula for partial and total regressions

Suppose that Y is a response, X an explanatory variable and that we are interested in the dependence of Y on X conditionally on U. Suppose further that U is unobserved. Suppose we are really interested in the dependence of Y on X, U jointly but can observe only dependence of Y on X.

An unobserved confounder



Above formula shows that $\beta_{YX.U} = \beta_{YX}$ if and only if

$$\beta_{YU.X}\beta_{UX} = 0,$$

Requiring either that U has no (linear) effect on Y once we have accounted for X or that U and X are unrelated. The second condition is satisfied if X is a randomized treatment (and U prior to X).

General distributions

By the quantile regression formula if $\gamma_{UX}(u, x) = 0$

$$\gamma_{YX}(y; x) = \int \gamma_{YX.U}(y; x, u) f_{U|YX}(u; y, x) du.$$

Various qualitative conclusions follow. Randomization preserves the primary features of the distribution of Y given both X and U in the conditional distribution given only X.

Multivariate response or outcome variables

Two broad possibilities

- components have an individual identity which should be preserved
- transformations of the components allowable to achieve clearer interpretation

Relatively simple case (*J. Mult. An.* **42** (1992), 162-170).

Not so simple time series case (*Proc. Nat. Acad. Sci* **96** (1999), 12273-12274).

Multivariate responses

Vector Y of response variables.

Two cases

- components individually interpretable
- at least for some interpretive purposes, transformation of components reasonable.

Simple formulation of 2. Vectors of responses Y and of explanatory variables X. Transform Y to $Y^* = AY$ so that Y_1^* depends only on X_1, etc. In simple 2×2 case leads to chordless four-cycle or seemingly-unrelated regression model.

Special case When $\dim(Y) = \dim(X)$ solution is

$$Y^* = \Sigma_{xx}\Sigma_{yx}^{-1}Y.$$

Note

$$\mathrm{cov}(Y^*, X) = \mathrm{cov}(X, X).$$

In general

$$Y^* = \Sigma_{xx}(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}Y.$$

An example

Preoperative patients

- Y_1, log palmitic acid
- Y_2, log linoleic acid
- Y_3, log oleic acid

- X_1 blood sugar
- X_2 sex

$$\hat{A} = \begin{pmatrix} 110.3 & 17.5 & -163.5 \\ -3.0 & 8.1 & -9.7 \end{pmatrix}$$

Simple interpretation

Time dependent variables

Suppose initially that Y is observed at two time points giving Y_2 and Y_1. For the moment ignore X. Matrix B_{21} of regression coefficients of Y_{2i} on $Y_{11}, \ldots, Y_{1p}$. Now transform both vectors by the same matrix A to give $Y^* = AY$. This gives a new matrix of regression coefficients

$$B^*_{21} = AB_{21}A^{-1}.$$

This is diagonal if and only if

$$AB_{21} = DA,$$

where D is diagonal. That is the rows of A are left eigenvectors of B_{21}.

Some complications

In the equation

$$AB_{21} = DA,$$

The matrix B_{21} is not in general symmetric.

Some consequences

- when B_{21} is replaced by an estimate $\hat{B}_{21}$ a significant imaginary component in particular to one of the leading eigenvalues would imply inconsistency with the formulation
- how would this be tested?
- essentially zero eigenvalues would have clear interpretation
- extension to more than two time points
- inclusion of X

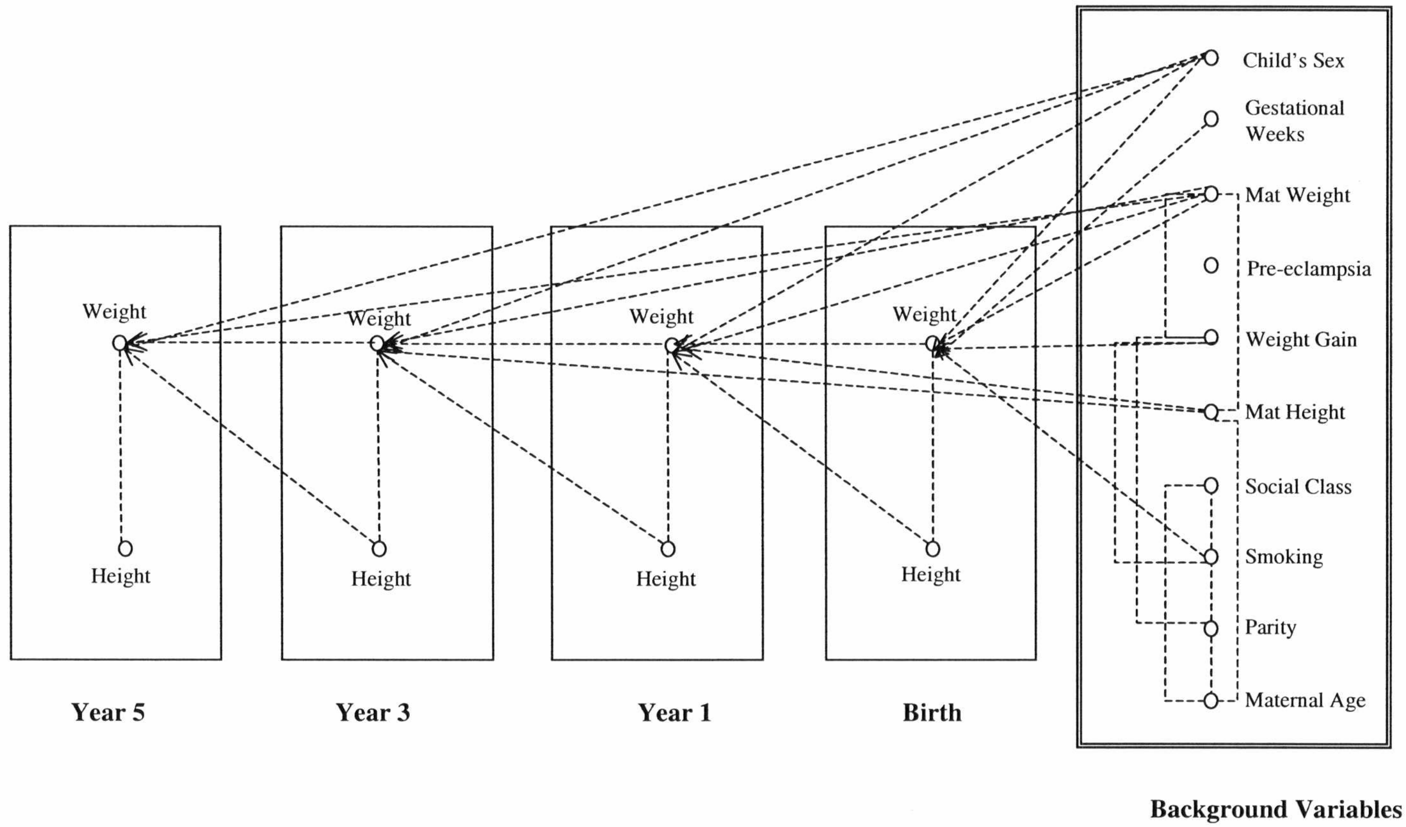
Cox, D.R. and Wermuth, N. (1999). Derived variables for longitudinal studies. *Proc. Nat. Acad. Sci.* **96**, 12273-12274.

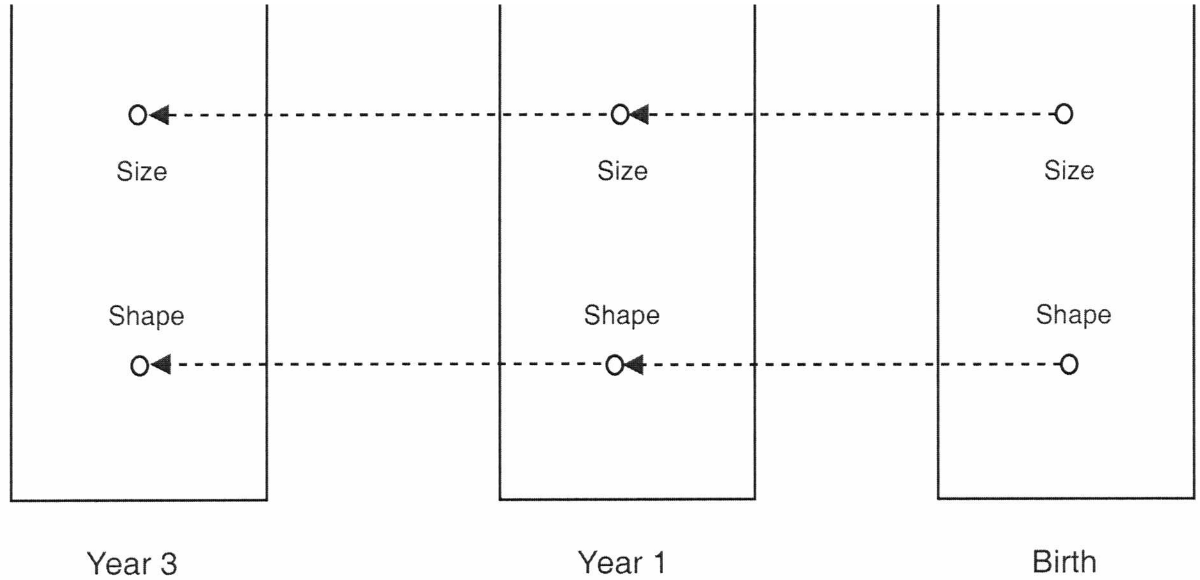
The Barry-Caerphilly milk study

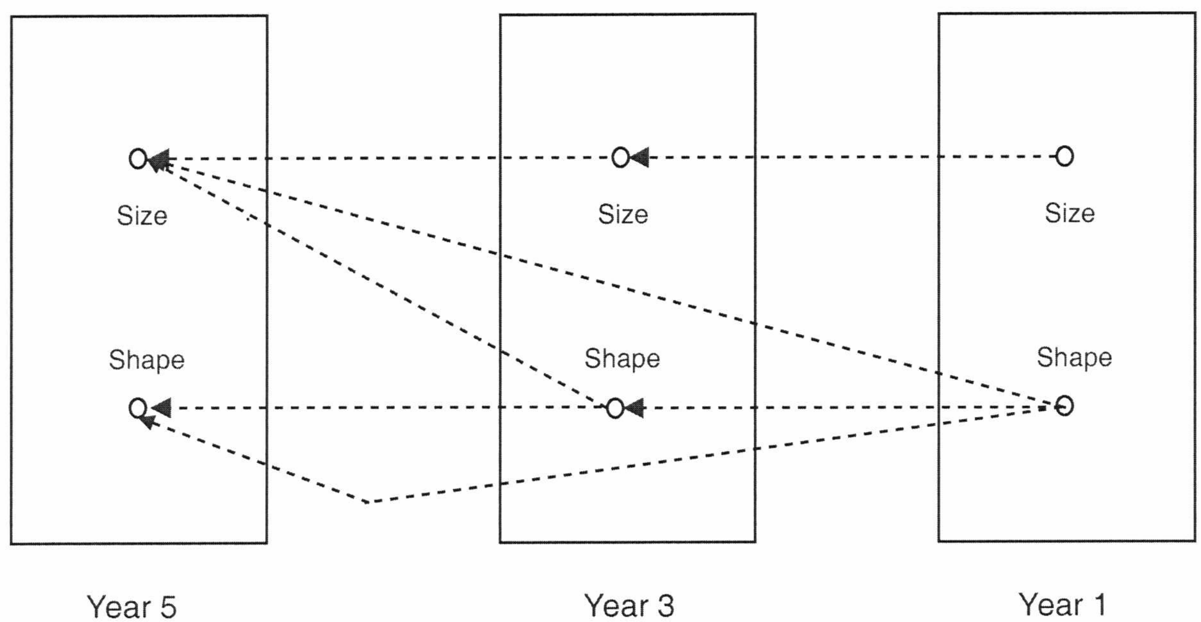
Outline of study

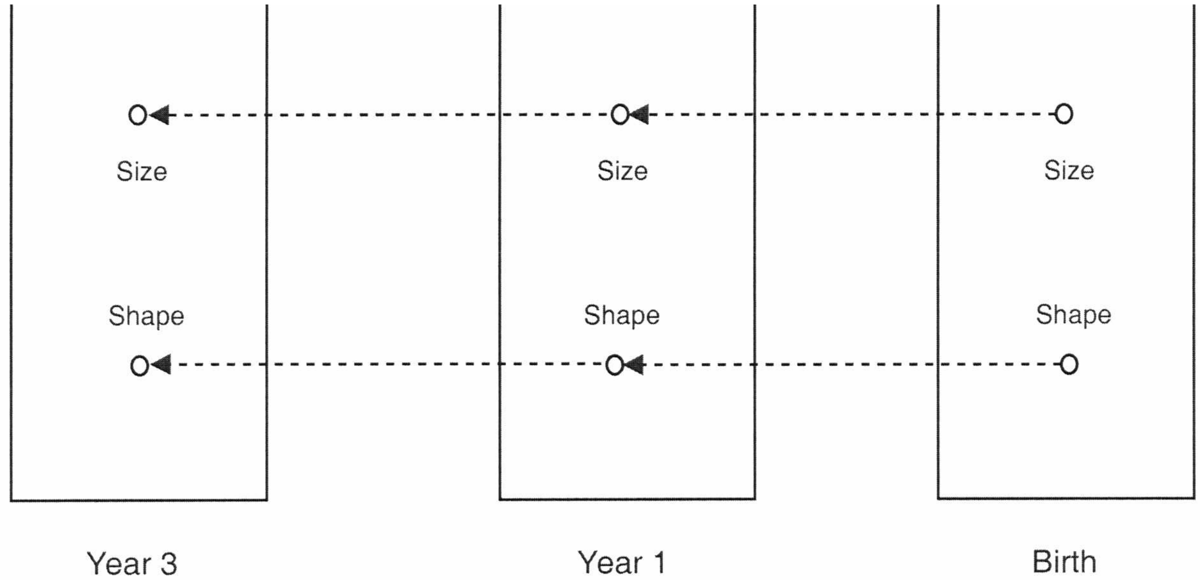
Work of Dr Andrew Roddam

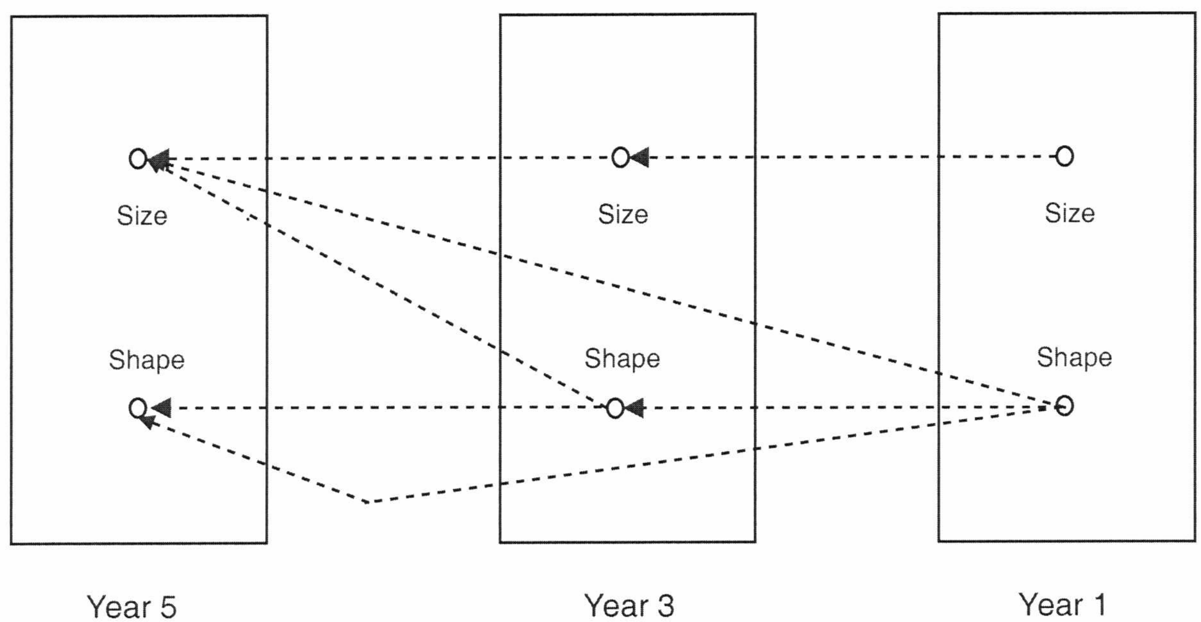
Figure 5.3: Fitted graphical model for the marginal analysis of the log weight of children.




Size
Size
Size
Shape
Shape
Shape
Year 3
Year 1
Birth

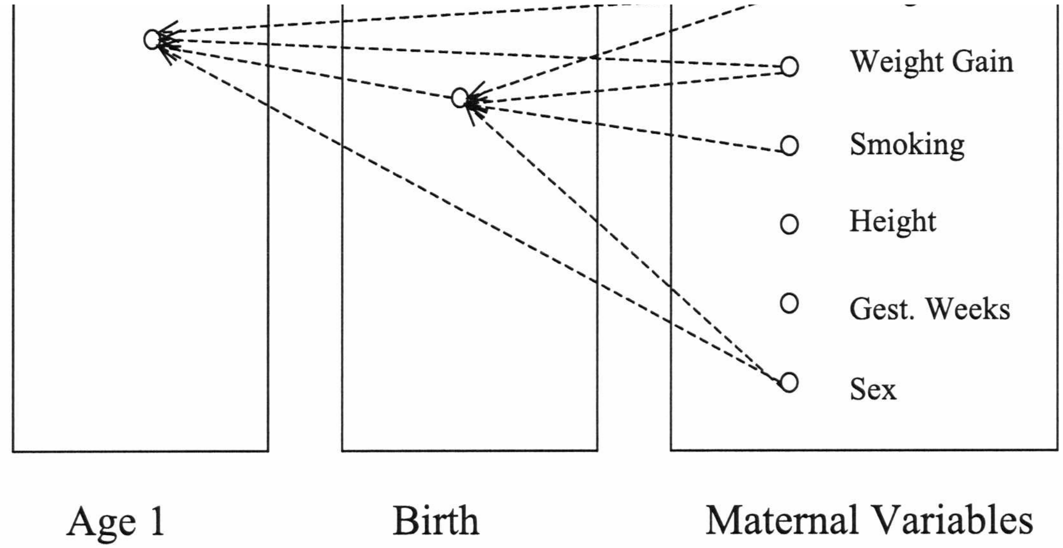

Size
Size
Size
Shape
Shape
Shape
Year 5
Year 3
Year 1


Size
Size
Size
Shape
Shape
Shape
Year 3
Year 1
Birth


Size
Size
Size
Shape
Shape
Shape
Year 5
Year 3
Year 1

Summary

- Structure of design
- Plan of analysis
- Detailed form of qualitative conclusions
- Presentation
- Derivation of associated properties



Evolution of Size Component