Int J Public Health (2010) 55:701–703
DOI 10.1007/s00038-010-0184-x

HINTS & KINKS

Using Directed Acyclic Graphs to detect limitations of traditional regression in longitudinal studies

Erica E. M. Moodie · D. A. Stephens

Received: 7 May 2010 / Revised: 17 June 2010 / Accepted: 16 July 2010 / Published online: 14 September 2010

Abstract
Introduction Longitudinal data are increasingly available to health researchers; these present challenges not encountered in cross-sectional data, not the least of which is the presence of time-varying confounding variables and intermediate effects.
Objectives We review confounding and mediation in a longitudinal setting and introduce causal graphs to explain the bias that arises from conventional analyses.
Conclusions When both time-varying confounding and mediation are present in the data, traditional regression models result in estimates of effect coefficients that are systematically incorrect, or biased. In a companion paper (Moodie and Stephens in Int J Publ Health, 2010b, this issue), we describe a class of models that yield unbiased estimates in a longitudinal setting.

Keywords Confounding · Mediation · Directed Acyclic Graphs · Longitudinal data

Directed Acyclic Graphs

Directed Acyclic Graphs (DAGs), also called *causal graphs*, formalize the causal assumptions that a researcher may make regarding the variables he wishes to analyze. A graph is said to be *directed* if all inter-variable relationships are connected by arrows indicating that one variable causes changes in another and *acyclic* if it has no closed loops (no feedback between variables)—see, for example, Greenland et al. (1999) or Pearl (2009). DAGs are becoming more common as researchers seek to clarify their assumptions about hypothesized relationships and thereby justify modeling choices (e.g., Bodnar et al. 2004; Brotman et al. 2008).

DAGs are particularly useful for visualizing and understanding confounding and mediation. A variable, L, is said to *confound* a relationship between an exposure, A, and an outcome, Y, if it is a common cause of both the exposure and the outcome, that is, if there is an arrow from L into A, and another from L into Y. If the effect of L on A and Y is not accounted for, it may appear that there is a relationship between A and Y when in fact their co-variation may be due entirely to changes in L. In cross-sectional data, blocking of a confounding effect is typically achieved by adjusting for the variable in a regression model.

L is said to *mediate* the effect of A on Y if it lies on the *causal pathway* between them, that is, there is at least one path from A to Y that passes through L. Specifically, if there is an arrow from A into L (so that exposure causes changes in this variable) and an arrow from L into the Y (indicating that the variable causally affects the outcome), L is called a *mediating variable*. To quantify the *total* effect of the exposure on the outcome, it is important not to block any effect that acts through a mediating variable. Therefore, we should not adjust for the mediating variable: including L in the regression model typically biases the effect estimate of A by underestimating the true value. Clearly, understanding the underlying structure of a dataset is required to make appropriate modelling decisions.

E. E. M. Moodie (✉)
Department of Epidemiology and Biostatistics,
McGill University, 1020 Pine Avenue West,
Montreal, QC H3A 1A2, Canada
e-mail: erica.moodie@mcgill.ca

D. A. Stephens
Department of Mathematics and Statistics,
McGill University, 805 Sherbrooke Street West,
Montreal, QC H3A 2K6, Canada

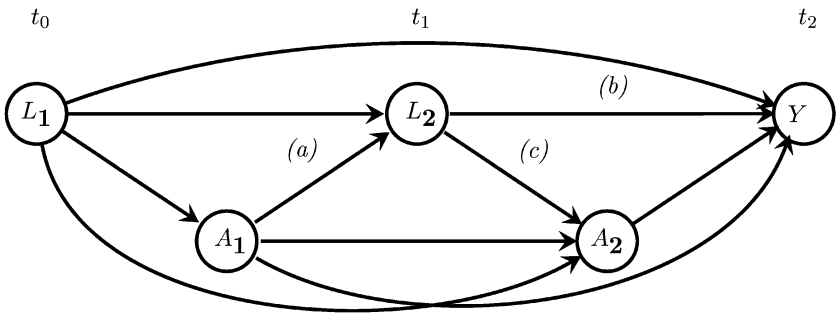

Fig. 1 A simple, two-interval Directed Acyclic Graphs. The (causal) relationship between A_1 and L_2, L_2 and Y, and A_2 and Y, are represented by arrows labeled *(a)*, *(b)*, and *(c)*, respectively. Time, denoted t_0, t_1, t_2, is marked above the variables

Table 1 Simulation of amblyopia study: data-generation protocol and summary of the sample study ($n = 100$)

Variable	Notation	Data generation	Summary
Baseline visual acuity	L_1	$E[L_1] = 0.1$	0.098
Occlusion, weeks 0–2	A_1	$P[A_1\|L_1] = \text{logit}^{-1}(-0.28 + 3.7L_1)$	54
Two-week visual acuity	L_2	$E[L_2\|A_1,L_1] = L_1 - 0.086A_1$	0.052
Occlusion, weeks 2–4	A_2	$P[A_2\|L_1,A_1,L_2] = \text{logit}^{-1}(-0.4 + 0.3A_1 + 4.5L_2)$	56
Four-week visual acuity	Y	$E[Y\|L_1,A_1,L_2,A_2] = -2L_1 + 0.22A_1 + 3L_2 - 0.086A_2$	0.031

Continuous variables summarized by mean, binary variables by count. All continuous variables were generated with a standard deviation of 0.02logMAR

Modelling choices become more complex when data are collected over time, particularly as a variable may act as both a confounder and a mediator. Consider a two-interval setting where data are collected at three times: baseline, t_1, and t_2. Potential confounding and/or mediating covariates are denoted L_1 and L_2, measured at baseline and t_1, respectively. Exposure in the intervals (t_0, t_1) and (t_1, t_2) is denoted A_1 and A_2, respectively. Outcome, measured at t_2, is denoted Y. See Fig. 1.

First, we first focus on the effect of A_1 on Y; A_1 acts directly on Y, but also acts indirectly through L_2 as indicated by arrows (*a*) and (*b*). As discussed above, it is important not to adjust for a variable that is on the causal pathway between exposure and outcome, so we must not adjust for L_2. We now turn our attention to the effect of A_2 on Y; L_2 confounds this relationship, as can be observed by arrows (*b*) and (*c*). To obtain an unbiased estimate of the effect of A_2, we must block on L_2. Thus, we have learned that to obtain unbiased estimators for both effects of interest, we must simultaneously adjust and not adjust for a time-varying covariate!

A simulated example

Suppose that researchers are interested in the effect of occlusion (eye-patching) on amblyopia in children, as in the Monitored Occlusion Treatment of Amblyopia Study (Stewart et al. 2002, 2004), which suggested that 100 h of occlusion improves visual acuity by −0.086logMAR (Moodie and Stephens 2010a) (a decrease on the logMAR scale corresponds to improved vision). The following simulated example is designed to follow the causal structure in Fig. 1, using a sample size of 100. Visual acuity is measured at baseline (L_1), 2 weeks (L_2), and 4 weeks (Y). Exposure A_1 is a binary indicator of occlusion compliance for at least 100 h between baseline and week 2, and A_2 the corresponding indicator for occlusion between weeks 2 and 4.

The data-generation protocol and simulated sample are described in Table 1. The data are analyzed using two traditional regression models:

$$\begin{aligned}\text{Model 1:}\quad & E[Y|A_1,A_2,L_1] = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 L_1,\\ \text{Model 2:}\quad & E[Y|A_1,A_2,L_1,L_2] = \beta_0 + \beta_1 A_1 + \beta_2 A_2\\ & \qquad + \beta_3 L_1 + \beta_4 L_2;\end{aligned}$$

results are presented in Table 2. Model 1 fails to account for the confounding by L_2 and hence yields a biased estimate of the effect of A_2. Model 1 inappropriately blocks the effect of A_1 on Y that acts through L_2, yielding a biased estimate of the effect of A_1.

To assure the reader that the results from a single data set of a modest size ($n = 100$) are not uncharacteristic of the problem, we repeated the simulation study 10,000 times. The bias of the estimator of the coefficient of A_2 in Model 1 was 6.0% and of the coefficient of A_1 in Model 2 was 679%. Repeating this using 10,000 datasets of size $n = 5{,}000$, the bias of A_2 in Model 1 was 6.3% and of A_1 in Model 2 was 679%, indicating that the bias does not diminish with increasing sample size. We conclude that the

Table 2 Simulation of amblyopia study: results from traditional regression models

Variable	Model 1			Model 2		
	$\hat{\beta}$	SE	% bias	$\hat{\beta}$	SE	% bias
A_1	−0.0356	0.0121	6.4	0.213	0.009	660.8
A_2	−0.0738	0.0121	14.2	−0.085	0.004	0.9

Model 1 adjusts for baseline visual acuity (L_1) only, while Model 2 adjusts for both baseline and 2-week visual acuity (L_1 and L_2). True parameter values are −0.038, and −0.086

example dataset produced typical results, and indeed traditional regression analyses provide biased estimators of the total effect of exposures in a longitudinal setting.

Discussion

We have presented here some of the challenges that arise in conventional modelling of time-varying exposures in longitudinal data, and how these can be detected through the formalization of assumptions using Directed Acyclic Graphs. While these results may be surprising, they should not be discouraging. As we explained in a companion article (Moodie and Stephens 2010b), there exists a class of models, the marginal structural models, that provide unbiased estimates in this setting.

Acknowledgments Both authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC). Moodie also acknowledges funding from the Canadian Institutes of Health Research (CIHR).

References

Bodnar LM, Davidian M, Siega-Riz AM, Tsiatis A (2004) Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. Am J Epidemiol 159:926–934

Brotman RM, Klebanoff MA, Nansel TR, Andrews WW, Schwebke JR, Zhang J, Yu KF, Zenilman JM, Scharfstein DO (2008) A longitudinal study of vaginal douching and bacterial vaginosis a marginal structural modeling analysis. Am J Epidemiol 168:188–196

Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10:37–48

Moodie EEM, Stephens DA (2010a) Estimation of dose-response functions for longitudinal data using the generalized propensity score. Stat Methods Med Res. doi:10.1177/0962280209340213

Moodie EEM, Stephens DA (2010b) Marginal structural models: unbiased estimation for longitudinal studies. Int J Publ Health (this issue)

Pearl J (2009) Causality, 2nd edn. Cambridge University Press, London

Stewart CE, Fielder AR, Stephens DA, Moseley MJ (2002) Design of the Monitored Occlusion Treatment of Amblyopia Study (MOTAS). Br J Ophthalmol 86:915–919

Stewart CE, Moseley MJ, Stephens DA, Fielder AR (2004) Treatment dose-response in amblyopia therapy: the Monitored Occlusion Treatment of Amblyopia Study (MOTAS). Investig Ophthalmol Vis Sci 45:3048–3054