Regression Based Causal Induction With Latent Variable Models

Lisa A. Ballesteros
Experimental Knowledge Systems Laboratory
University of Massachusetts/Amherst
Box 34610
Amherst, MA 01003-4610
balleste@cs.umass.edu
(413) 545-3616

Scientists largely explain observations by inferring causal relationships among measured variables. Many algorithms with various theoretical foundations have been developed for causal induction e.g., (Spirtes, Glymour, & Scheines 1993; Pearl & Verma 1991), but it is widely believed that regression is ill-suited to the task of causal induction. Multiple regression techniques attempt to estimate the influence that regressors have on a dependent variable using the standardized regression coefficient, β. Assuming the relationship among variables is linear, β_{YX} measures the expected change in Y produced by a unit change in X with all other predictor variables held constant.

Arguments against using regression methods for causal induction rest on the fact that the error in estimating β_{YX} can be large, particularly when unmeasured or latent variables account for the relationship between X and Y, or when X is a common cause of Y and another predictor (Mosteller & Tukey 1977; Spirtes, Glymour, & Scheines 1993). In fact, β may suggest X has a strong influence on Y when it has little or none. We have developed a regression-based causal induction algorithm called FBD (Cohen *et al.* 1994) which performs well in these situations.

The heuristic that is primarily responsible for making FBD less sensitive to the above problems is the ω score. Let r_{XY} be the correlation between X and Y, and $\omega = (r_{YX} - \beta_{YX})/r_{YX}$. ω measures the proportion of r_{YX} not due to the direct effect of X on Y. If ω_{YX} exceeds a threshold, X is pruned from the set of candidate predictors. This threshold is set arbitrarily by the user, but we are exploring the use of clustering algorithms to set it by partitioning the ω values of the predictor variables.

Spirtes et al. describe four causal models (1993, p. 240) for which their studies showed regression methods performed poorly by always choosing predictors whose relationship to the dependent variable is mediated by latent variables or common causes. One model is reproduced in Figure 1. The difficulty with this model is that the error in the estimate for β_{X_2Y} may be large due to X_2's relationship to X_3 via the latent variable T_1.

To determine the susceptibility of FBD to latent variable effects, we tested the performance of FBD[1] on latent variable models, and ran stepwise regressions as a control. Twelve sets of coefficients for the structural equations for each of Spirtes's models were generated, as were data sets for each, having 100 to 1000 variates. Each sample was given to FBD and to MINITAB's stepwise procedure. Performance was measured by the number of times the algorithm incorrectly chose predictors related via latent variables and the number of times it chose correctly. FBD chose 88% of the correct predictors, while MINITAB chose 93% of them. On the other hand, FBD rejected variables whose relationships to the dependent variable were due to latent or common causes 82% of the time, while MINITAB rejected them only 25% of the time. Thirty nine percent of FBD's rejections were due to ω. Although MINITAB got a slightly higher hit rate for correct predictors than did FBD, FBD got fewer false positives. These results suggest that ω makes FBD less susceptible to latent variable effects than standard regression techniques. FBD's ability to avoid the problems described above make it a promisng causal induction algorithm.

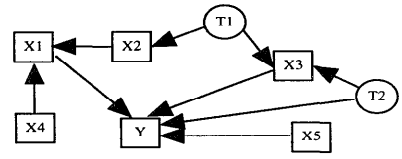

Figure 1: Latent Variable Model

[1] In comparison studies among FBD, Pearl's IC, and Spirtes's PC, FBD performed better on all of our measures of performance (Cohen *et al.* 1994; Gregory & Cohen 1994).

References

Cohen, P. R.; Ballesteros, L.; Gregory, D.; and St.Amant, R. 1994. Regression can build predictive causal models. Submitted to the Tenth Annual Conference on Uncertainty in AI. Technical Report 94-15, Dept. of Computer Science, University of Massachusetts/Amherst.

Gregory, D., and Cohen, P. R. 1994. Two algorithms for inducing causal models from data. Submitted to Knowledge Discovery in Databases Workshop, Twelfth National Conference on Artificial Intelligence.

Mosteller, F., and Tukey, J. W. 1977. *Data Analysis and Regression, A Second Course in Statistics.* Addison-Wesley Publishing Company.

Pearl, J., and Verma, T. 1991. A statistical semantics for causation. *Statistics and Computing* 2:91–95.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search.* Springer-Verlag.