Regression Based Causal Induction With Latent Variable Models

Lisa A. Ballesteros
Computer Science Department, LGRC
University of Massachusetts/Amherst
balleste@cs.umass.edu

Scientists have largely explained observations by inferring causal relationships between measured variables. These relationships are tested experimentally when possible and can be used as guidance for further experimentation, and as building blocks for causal theories and models. Determining the true nature of these relationships can be difficult. For example, given two related variables X and Y, there are three possible explanations for their relationship: X may cause Y, Y may cause X, or the two may be correlated due to their relationship to a third variable and thus have no causal relationship at all. Another difficulty is that the number of possibilities grows exponentially with the number of variables in the dataset so the number of possible causal models for n variables is 2^{n^2-n}. If we are going to search for a model, search control heuristics are necessary to prune the number of models searched in order to keep the size of the search space from being prohibitive. We also want a causal model which reflects the relationships existing in a dataset.

Many algorithms with various theoretical foundations have been developed for causal induction [7,6,2]. Multiple regression techniques attempt to estimate the influence that regressors have on the dependent variable using the standardized regression coefficient, β_{YX}. Assuming the relationship among the variables is linear, the parameter β_{YX} measures the expected change in Y produced by a unit change in X with all other predictor variables held constant. Regression models include variables for which β is large. Descriptions of regression methods can be found in any standard regression text [3].

It is widely believed that regression is ill-suited to the task of causal induction. Arguments against using regression methods rest on the fact that the error in estimating β_{YX} can be quite 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 [5,7]. 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 [1] which performs well in these situations.

FBD uses several heuristics to prune variables which are judged to be poor predictors. The heuristic that makes 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} that may be due to the direct effect of X on Y. If ω_{YX} falls below a threshold, X is pruned from the set of predictors. This threshold is set arbitrarily by the user, but we are currently exploring the use of clustering algorithms to set it by partitioning the ω values of the predictor set. We have not formally shown why ω is a good pruning heuristic, but will give the intuition behind it. The correlation coefficient, r_{YX}, measures the linear relationship between X and Y. This relationship may be composed of a direct effect plus indirect effects mediated by X's and Y's relationships to other variables. Recall that β_{YX} estimates the direct effect of X on Y. A high ω value suggests that the correlation between two variables is largely due to X's and Y's relationships to other variables, therefore we do not gain predictive power by choosing X as a predictor of Y.

FBD has performed well in experiments designed to test the accuracy with which a causal induction algorithm is able to reconstruct causal structure from statistical data[1]. We constructed sixty causal models having six, nine, or twelve, twenty models of each size. Artificial data sets were then generated for each model using its linear equations. FBD was run on the data sets and its performance was evaluated by a number of criteria which measure the similarity of the model generated by the algorithm to the model which generated the data set i.e., the target model. The four evaluation functions which we believe are most important are as follows: $DependentR^2$ measures the proportion of the variance in the dependent variable accounted for by its predictors; $\Delta\ R^2$ is the average difference between the R^2 for a predicted variable in the target and the R^2 for that same node in the model produced by the algorithm, and measures FBD's ability to predict those variables predicted by the target model; correct% measures the percentage of direct links in the algorithm's model which are also in the target model; and wrong/correct gives the number of incorrect links for each correct link. Results show that FBD builds models which account for a large proportion of the variance in predicted variables, recall is at about 67%, and gets .76, 1.18, and 1.66 wrong links for every correct link for the 6, 9, and 12 variable models respectively. In comparison studies, FBD[2] performed better on all these measures than did either Pearl's IC algorithm or Spirtes'

[1] We have run comparison studies [1,4] with FBD and Pearl's and Spirtes' causal induction algorithms, but space precludes discussing them here.

[2] FBD requires knowledge of which variable is the dependent variable. We have developed the FTC algorithm, based on FBD, which performs as well or better with respect to some measures and requires no knowledge of the dependent variable.

PC algorithm.

Spirtes et al. describe four causal models [7, p. 240] for which their studies showed regression methods performed poorly by always choosing predictors whose relationship to the dependent is mediated by latent variables or common causes. One of the models is reproduced in Figure 1. The difficulty that arises 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 which is mediated by the latent variable T_1.

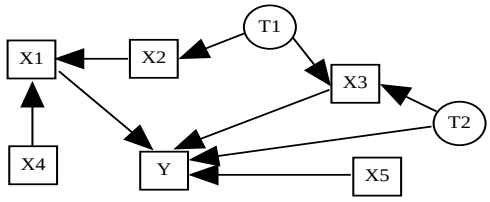

Figure 1: Latent Variable Model

To test FBD's susceptibility to be led by high errors in the estimate of β's to choose predictors incorrectly, we ran experiments in which we compared the performance of FBD to that of stepwise regression. Twelve different sets of coefficients for the structural equations for each of Spirtes' models were generated, as were data sets for each having 100 to 1000 variates. Each sample was given to FBD and to MINITAB's[3] stepwise regression procedure (.05 significance level). Performance on these data was measured by the number of times the algorithm chose the correct predictors and number of times it chose incorrectly. FBD chose the correct predictor(s) 88% of the time, while MINITAB chose correctly 93% of the time. Variables from the set of correct predictors were always included in the predictor set, the remaining 12% and 7% were due to not choosing the entire set of correct predictors. However, when the algorithms could have incorrectly chosen variables whose relationships with the dependent variable were due to latent variables or common causes, FBD rejected them 82% of the time, while MINITAB rejected them only 25% of the time. 39% of FBD's rejections were due to omega pruning. Although MINITAB got a slightly higher hit rate for correct predictors than did FBD, FBD got fewer false positives than did the stepwise procedure. These results suggest that omega pruning makes FBD less susceptible to latent variable effects than standard regression techniques.

References

[1] P. R. Cohen, L. Ballesteros, D. Gregory, and R. St. Amant. Regression can build predictive causal models. Submitted to the Twelfth National Conference of the American Association for Artificial Intelligence. Technical Report 94-15, Dept. of Computer Science, University of Massachusetts/Amherst, 1994.

[2] G. Cooper and E. Herskovits. A bayesian method for constructing bayesian belief networks from databases. In Bruce D'Ambrosio, Philippe Smets, and Piero Bonissone, editors, *Uncertainty in Artificial Intelligence*, pages 86–94. Morgan Kaufmann, 1991.

[3] N.R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley and Sons, 1966.

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

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

[6] Judea Pearl and T.S. Verma. A statistical semantics for causation. *Statistics and Computing*, 2:91–95, 1991.

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

[3] MINITAB is a commercial, interactive statistical package