2000

Estimating Latent Causal Influences: TETRAD III Variable Selection and Bayesian Parameter Estimation

Richard Scheines
Carnegie Mellon University

Estimating Latent Causal Influences:
TETRAD III Variable Selection and Bayesian Parameter Estimation[1]

Richard Scheines
Department of Philosophy, Carnegie Mellon University

Abstract

The statistical evidence for the detrimental effect of exposure to low levels of lead on the cognitive capacities of children has been debated for several decades. In this paper I describe how two techniques from artificial intelligence and statistics help make the statistical evidence for the accepted epidemiological conclusion seem decisive. The first is a variable-selection routine in TETRAD III for finding causes, and the second a Bayesian estimation of the parameter reflecting the causal influence of Actual Lead Exposure, a latent variable, on the measured IQ score of middle class suburban children.

1. Introduction.

This paper presents an example of causal discovery in which two pieces of artificial intelligence technology proved crucial. The pieces are TETRAD III's Build module, used to identify and discard spurious confounders of the relationship between lead exposure and IQ, and TETRAD III's Gibb's sampler[2] used to estimate the influence of lead exposure on IQ within an underidentified model involving the remaining non-spurious confounders.

In a variety of contexts in KDD, effective variable selection is crucial. For example, identifying a small set of variables that can accurately predict who will be profitable as a credit card customer is a classic KDD problem in the financial realm. Variable selection techniques accompany different models. For example, one can use best subsets, or a

[1] To appear in Handbook of Data Mining (2001), MIT Press.
[2] See: http://hss.cmu.edu/philosophy/TETRAD/tetrad.html

forward or backward stepwise procedure with linear or logistic regression. Decision trees all use some variant of a forwards stepwise procedure. In each case, the typical desiderata is purely predictive. That is, the goal is to use observations on some variables in order to predict the values of other variables. In contrast, another goal for KDD can be causal discovery. Instead of using a given data set to build a model which will then be used with future data to predict one variable from the *observations* on the model's inputs, the goal is to use a given data set to build a model which will then be used to predict one variable from *interventions* that *set* the values of the model's inputs. Mining a data set might tell us that the amount of stain on a person's teeth is a good predictor of how much tea they will buy, but it does not tell us that a person will buy more tea if we first sell them gum that stains their teeth. In predicting the effects of interventions, causal knowledge is essential, and standard variable selection techniques simply are not designed to identify the causes of a variable. The TETRAD programs (Scheines, et al., 1994) implement techniques that are designed to identify the causes and not just the predictors of a variable. In this case study, the difference between standard variable selection in regression and in TETRAD III are made vivid.

In the paper that follows, I first briefly survey the history of lead and IQ research that led to this work. I then discuss variable selection for prediction vs. variable selection for causal discovery, and illustrate the difference on this case. In the final sections, I show how these results allowed a Bayesian estimation of the final causal model to give persuavive evidence that exposure to lead does indeed have a deleterious effect on cognition, even at low levels.

2. A Brief History of Lead and IQ Research

By measuring the concentration of lead in a child's baby teeth, Herbert Needleman was the first epidemiologist to reliably measure cumulative lead exposure in children. His work helped convince the United States to eliminate lead from gasoline and most paint (Needleman, et. al., 1979). Needleman and a few colleagues collected data[3] on over 50 variables on almost 300 New England children. The measures included the child's IQ, parental education, parental IQ, SES, teacher evaluations of the child, response time tests,

several other possible confounders, and finally a measure of cumulative lead exposure. The goal was causal: to estimate the effect of cumulative lead exposure on the child's IQ.

Needleman's original statistical analysis, which was basically ANOVA (Needleman, et. al., 1979) was criticized by the EPA (Grant, et al., 1983), who concluded that his data neither supported nor rejected the conclusion that lead was damaging at the doses he recorded in asymptomatic children. Needleman reanalyzed his data with multiple regression. He performed a variable selection search using backwards stepwise eliminative regression, and found five covariates (measured confounders) that he included in a multiple regression to estimate the effect of lead on IQ. He found that even after controlling for the five covariates, the estimated effect of lead on IQ was negative and significant (Needleman, et al., 1985).

This helped with the EPA, but aroused other worries from Steve Klepper, an economist at Carnegie Mellon (see Klepper, 1988; Klepper, Kamlet, & Frank, 1993). Klepper correctly argued that Needleman's statistical model (a linear regression) neglected to account for measurement error in the regressors. That is, Needleman's measured regressors were in fact imperfect proxies for the actual but latent causes of variations in IQ, and in these circumstances a regression analysis gives a biased estimate of the desired causal coefficients and their standard errors.

Unfortunately, an errors-in-all-variables model that explicitly accounts for measurement error is "underidentified," and thus cannot be estimated by classical techniques without making additional assumptions. Klepper, however, had worked out an ingenious technique to bound the estimates, provided one could reasonably bound the amount of measurement error contaminating other measured covariates (Klepper, 1988, 1993). The bounds on the measurement error infecting the measured covariates required to estimate the effect of actual lead exposure in Needleman's model seemed unreasonable, however, and Klepper concluded that the statistical evidence for Needleman's hypothesis was indecisive.

Reanalyzing Needleman's data, I used TETRAD III to check whether backwards stepwise regression had indeed identified the appropriate set of confounders that should

[3] The data are available in the Datasets section of Statlib at Carnegie Mellon: www.statlib.cmu.edu.

be included in the final model. TETRAD III discarded three of the five covariates that stepwise regression had located, and these variables were precisely the ones which required unreasonable measurement error assumptions in Klepper's analysis. With the remaining regressors, I specified an errors-in-all-variables model to parameterize the effect of actual lead exposure on children' IQ. This model is still underidentified, but instead of trying to bound the parameters of interest I put a prior distribution over the parameters in the model and used a Gibbs sampler (Smith and Roberts, 1993, Scheines, Hoijtink, and Boomsma, 1999) to do a Bayesian estimation of the resulting model. Under several priors, nearly all the mass in the posterior was over negative values for the effect of actual lead exposure--now a latent variable--on measured IQ.

3. Variable Selection with TETRAD III

In their 1985 article in *Science*, Needleman, Geiger and Frank gave results for a multivariate linear regression of children's IQ on lead exposure. Having started their analysis with almost 40 covariates, they were faced with a variable selection problem to which they applied backwards stepwise regression, arriving at a final regression equation involving lead and five covariates. The covariates were measures of genetic contributions to the child's IQ (proxied by the parent's IQ), the amount of environmental stimulation in the child's early environment (proxied by the mother's education), physical factors that might compromise the child's cognitive endowment (proxied by the number of previous live births and the parent's age at the birth of the child). The measured variables they used are as follows, with the correlations among these variables and the p-value of each correlation given in Table 1.

Ciq - child's verbal IQ score
Lead - measured concentration in baby teeth
Mab - mother's age at birth
Fab - father's age at birth
Med - mother's level of education in years
Nlb - number of live births previous to the sampled child
Piq - parent's IQ scores

Table 1. Correlations & p-values (n=221)

Correlations

	Lead	*Fab*	*Nlb*	*Med*	*Mab*	*Piq*	*Ciq*
Lead	1.00						
Fab	-.08	1.00					
Nlb	.11	.39	1.00				
Med	-.14	.02	-.18	1.00			
Mab	-.15	.85	.47	.003	1.00		
Piq	-.06	.17	.03	.53	.16	1.00	
Ciq	-.23	-.0003	-.17	.41	.05	.40	1.00

p-values

	Lead	*Fab*	*Nlb*	*Med*	*Mab*	*Piq*
Fab	.23					
Nlb	.10	.00				
Med	.04	.78	.01			
Mab	.02	.00	.00	.96		
Piq	.39	.01	.70	.00	.02	
Ciq	.00	.99	.01	.00	.43	.00

The standardized regression solution is as follows, with t-ratios in parentheses. Except for *Fab*, which is significant at 0.1, all coefficients are significant at 0.05, and $R^2 = .271$.

$$\hat{Ciq} = -.143\,Lead + .219Med + .247Piq + .237Mab - .204Fab - .159Nlb \qquad [1]$$
$$\qquad\quad (2.32) \qquad (3.08) \qquad (3.87) \qquad (1.97) \qquad (1.79) \qquad (2.30)$$

The intuition behind statistically "controlling" for covariates in a multivariate regression intended to estimate causal influence is scientifically appealing but can be wrong. It stems from the following plausible story: an association between *X* and *Y* might not be due to a direct causal link from *X* to *Y*, but rather at least partly due to confounders (common causes of *X* and *Y*), or intermediate causes; statistically controlling for covariates can remove that part of the association produced by confounders, leaving only the association between *X* and *Y* due to an actual causal relationship. In the case of linear regression, β_i (the regression coefficient of the outcome *Y* on X_i) is statistically

significant just in case the partial correlation of Y and X_i controlling for *all* of the other regressors is significant.

Measuring whether there is an association between X and Y after controlling for *all* the other potential confounders is the right test for whether X is a direct predictor of Y, but it is not necessarily the right test for whether X is a direct cause of Y. Clearly Needleman (and Klepper after him) considered the variable selection problem settled by the significance test for coefficients in the multivariate regression, and this seems to be standard operating procedure in the social science and epidemiological community. Unfortunately, the general principle is wrong, and this data set is an exemplar of why.

In the general setting of multivariate regression, linear or otherwise, an outcome Y and a set of regressors $\boldsymbol{X}$ is specified. Assuming that $\boldsymbol{X}$ is prior to Y, in which case Y cannot cause any $X \in \boldsymbol{X}$, we say that X is causally adjacent to Y relative to the set $\boldsymbol{X}$ just in case either X is a direct cause of Y relative to $\boldsymbol{X}$, or there is a Z not in $\boldsymbol{X}$ such that Z is a common cause of X and Y. TETRAD III requires two assumptions in order to analyze whether any $X_i \in \boldsymbol{X}$ is causally adjacent to Y relative to $\boldsymbol{X}$ from population data: the Causal Markov Condition and Faithfulness.[4] The Causal Markov Condition amounts to assuming that every variable X is independent of all variables that are not its effects conditional on its immediate causes (Spirtes, et al., 1993). The Causal Markov Condition is satisfied necessarily by structural equation models with independent errors (Kiiveri and Speed, 1982), and seems to be relatively uncontroversial. Faithfulness amounts to assuming that all independences true in a population determined by a causal structure are due to the absence of causal connection and not due to parameter values that produce independences by perfect cancellation. Although versions of this assumption are used in every science (Spirtes et al., 1993), it is not uncontroversial, and has been generally challenged by Robins and Wasserman (1996). Allowing these two assumptions, it turns out that X is causally adjacent to Y only if X and Y are dependent conditional on *every subset* of $\boldsymbol{X}$ - $\{X,Y\}$ (Spirtes, et al., 1993). Contrast this criterion with the one used in multivariate regression: X is causally adjacent to Y only if X and Y are dependent

[4]For discussions of the reliability of regression for determing causal structure, see (Spirtes, et al., 1993, ch. 8; Scheines, 1995; and Glymour et al., 1994).

conditional on *exactly the set* $\boldsymbol{X}$ - $\{X_i, Y\}$. The model in Figure 1, in which $\boldsymbol{X} = \{X_1, X_2, X_3\}$ and Z is unmeasured, makes the flaw in the regression criterion vivid.

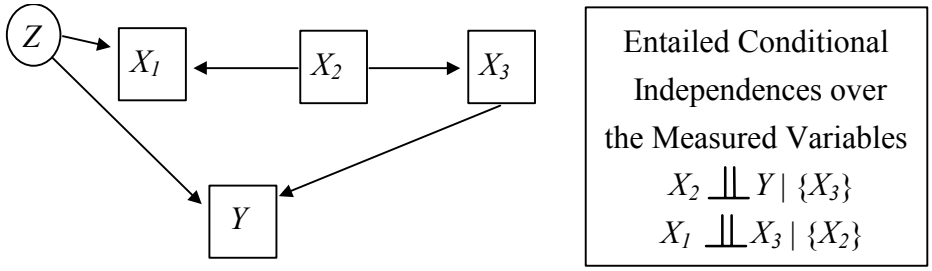

Figure 1: A model that fools regression

This model does not entail that X_2 and Y are independent when we condition on all the other regressors $\{X_1, X_3\}$. It is possible for the model to imply this independence, but only for unfaithful parameterizations. For all faithful parameterizations, a regression of Y on X will produce non-zero coefficients for all three regressors. Although this is *not* a sampling problem, it is easy to verify that regression will mistakenly conclude that X_2 is causally adjacent to Y on sample data by randomly parameterizing this model, generating a pseudo-random sample, and then running a regression of Y on X_1, X_2, and X_3.

It turns out that the regression criterion is reliable for causal adjacency only when X is known to be prior to Y *and* the measured variables are known to be **confounder complete**,[5] i.e., all common causes of two variables in $\boldsymbol{X} \cup \{Y\}$ are already in $\boldsymbol{X} \cup \{Y\}$. Assuming confounder completeness in general seems entirely unrealistic, and clearly so for the lead data.

The FCI algorithm executed by the Build module in TETRAD III does not assume confounder completeness, and asymptotically dominates regression as a test for causal adjacency. That is, with correct statistical decisions about independence, the FCI algorithm can detect non-causally adjacent variables that regression cannot, but not vice versa (Spirtes, Glymour, & Scheines, 1993). Run on the correlations in Table 1, TETRAD III indicates that only *Lead*, *Med*, and *Piq* are adjacent to *Ciq*, and that *Mab*, *Fab*, and *Nlb* are *not* causally adjacent to *Ciq*, contrary to the regression analysis. In

[5] In TETRAD III, and many previous publications, we use the terminology of "causal sufficiency" to mean what I define here as confounder completeness.

Needleman's data, *Mab*, *Fab*, and *Nlb* are more correlated with *Ciq* after conditioning on the other regressors than they are unconditionally. *Mab* and *Fab*, for example, are completely uncorrelated with *Ciq* unconditionally (see Table 1), yet are *correlated* with *Ciq* conditional upon all the other regressors. Whether *Mab* and *Fab* are measured with error or not, then under these assumptions they or the variables they are proxies for cannot be causally adjacent to *Ciq* relative to this set. The regressor *Nlb* is correlated with *Ciq* unconditionally, almost uncorrelated with *Ciq* when conditioned on *Med* ($r_{Nlb,Ciq.Med}$ = -.114, p = .1), but once again correlated when conditioned on the entire set of regressors.

To finalize the variable selection phase, I did a regression of *Ciq* on only those regressors found to be causally adjacent to *Ciq*, namely lead, *Med*, and *Piq*.

$$\hat{Ciq} = \underset{(2.89)}{-.177\ Lead} + \underset{(3.5)}{.251\ Med} + \underset{(3.59)}{.253\ Piq} \qquad [2]$$

The overall R^2 for the regression in equation [2] is .243, which is quite close to the R^2 of .271 from the full regression on all six variables in equation [1]. All coefficients in [2] are significant at .01, as expected, and the coefficient on lead is slightly more negative than it was in equation [1].

4. Estimating the Parameters of an "Underidentified" Model

As Klepper (1988, 1993) correctly points out, these measured regressor variables are really proxies that almost surely involve substantial measurement error. Measured lead is really a proxy for actual lead exposure, *Med* is really a proxy for environmental stimulation, and *Piq* is really a proxy for genetic factors related to IQ. Figure 2 shows a full errors-in-all variables specification for the variables included by TETRAD II.[6] The task is now to estimate the coefficient β_1.

[6] In this figure, measured variables are boxed, latent variables are enclosed in ovals, and error terms are left unenclosed.

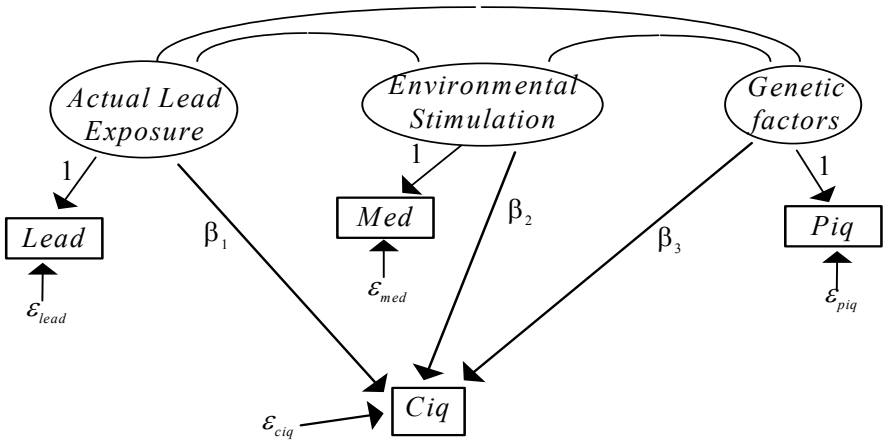

Figure 2. Errors-in-all-variables model for Lead's influence in IQ.

Although an errors-in-all-variables linear structural equation model seems a reasonable specification, this model is underidentified in the classical setting. That is, for any implied covariance matrix $\Sigma(\theta)$ that minimizes a discrepancy function of the implied and observed covariances, there are an infinity of parameterizations θ' such that $\Sigma(\theta) = \Sigma(\theta')$. In this case there are 13 free parameters in the model but only 10 data points in the covariance matrix for *Ciq*, lead, *Med*, and *Piq*, thus the model is underidentified by three degrees of freedom.

Several strategies exist for identifying the model. One is to specify the exact proportion of measurement error for each measured independent variable. Since in this model we have standardized the variables, $\sigma^2(Lead) = 1$. By the model's specification, $\sigma^2(Lead) = \sigma^2(Actual\ Lead) + \sigma^2(\varepsilon_{Lead})$, so the proportion of measured *Lead's* variance that is due to measurement error is just $\sigma^2(\varepsilon_{Lead})$, which is between 0 and 1. Similarly for the other regressors. Using a linear regression to estimate β_1 is equivalent to specifying a measurement error equal to zero for each regressor. We could also simply stipulate that the measurement error for lead is 0.20, or some other number.

Klepper and Leamer (1984) showed that in certain circumstances one could, by imposing bounds on how much measurement error is present, sometimes bound the actual coefficients in the underidentified errors-in-variables model. In 1988 and again in 1993 Klepper argued that the upper bounds required by his method to bound the true

coefficients in the lead errors-in-variables model (with all five covariates) were unreasonable. For example, one had to bound the measurement error for *Fab* (father's age at birth) at approximately 5%, which did not seem justifiable, considering *Fab* is a proxy for physical, emotional, and intellectual factors present in the father that might influence a child's IQ score. Performing Klepper's analysis on the reduced set of regressors identified by TETRAD II, one must be willing to bound the measurement of *Lead*, *Med*, and *Piq* at .710, .465, and .457 respectively, a combination of bounds of which I am reasonably confident. Klepper's technique, however, provides sufficient but not necessary conditions for bounding, and it cannot provide point estimates or standard errors.

The alternative I favor is Bayesian. By putting a prior distribution over the parameters and then computing the posterior, one can compute point estimates, e.g., the mean or median in the posterior (θ_{EAP} and θ_{MDAP}), standard deviations around the point estimates ($\sigma(\theta_{EAP})$), percentiles that can be used to compute posterior credibility intervals ($\theta_{.025}$ and $\theta_{.975}$) and many other statistics of interest. If the posterior cannot be computed analytically, which is certainly the case for all but the most trivial structural equation models, then one can now compute a sample from the posterior by MCMC simulation methods with TETRAD III (Scheines, Hoijtink, and Boomsma, 1999).[7] One can then use the sample from the posterior to estimate the posterior statistics from their sample counterparts, i.e., $\hat{\theta}_{EAP}$, $\hat{\theta}_{MDAP}$, $s(\hat{\theta}_{EAP})$, $\hat{\theta}_{.025}$, and $\hat{\theta}_{.975}$. For simplicity, I use a multivariate normal prior over the t parameters, i.e., $p(\boldsymbol{\theta}) \sim N_t(\boldsymbol{\mu}_0, \boldsymbol{\sigma}^2_0)$, and I enforce bounds on the parameters, e.g., variances are bounded below by 0, by rejecting sampled values outside of the legal parameter bounds.[8]

To apply the Bayesian solution to the lead problem, we must put a prior over the parameters. Needleman pioneered a technique of estimating cumulative lead exposure by measuring the accumulated lead in a child's baby teeth. Needleman guesses that between 0% and 40% of the variance in his measure of dentine lead is from measurement error, with 20% a conservative best guess. For the measures of environmental stimulation and

[7] A Gibbs sampler for computing the posterior over the parameters of a structural equation model is now available in TETRAD III: http://hss.cmu.edu/philosophy/TETRAD/tetrad.html

genetic factors, he was less confident, we guessed that between 0% and 60% of the variance in *Med* and *Piq* is from measurement error, with 30% as our best guess. Thus I began the Bayesian analysis by specifying a multivariate normal prior over the model's 13 parameters. The part of the prior involving measurement error is given in Table 2. The prior is otherwise uninformative.

Table 2. Multivariate Normal prior distribution over the measurement error parameters in the errors-in-all-variables model.

Parameter	Mean (μ_0)	Standard Deviation (σ_0)
$\sigma^2(\varepsilon_{Lead})$	0.200	0.10
$\sigma^2(\varepsilon_{Med})$	0.300	0.15
$\sigma^2(\varepsilon_{Piq})$	0.300	0.15

Using this partially informative prior, I produced 50,000 iterations with the Gibbs sampler in TETRAD III. The sequence converged immediately. Table 3 shows the results of this run, and the histogram in Figure 3 shows the shape of the marginal posterior over β_1, the crucial coefficient representing the influence of actual lead exposure on children's IQ. The results support Needleman's original conclusion, but do not require unrealistic assumptions about the complete absence of measurement error, or assumptions about exactly how much measurement error is present, or assumptions about upper bounds on the measurement error for the remaining regressors.

Table 3. Gibbs sample statistics for the causal parameters in the errors-in-all-variables model.

	$\hat{\theta}_{EAP}$	$\hat{\theta}_{MDAP}$	$s(\hat{\theta}_{EAP})$	$\hat{\theta}_{.025}$	$\hat{\theta}_{.975}$
β_1	-0.215	-0.211	0.097	-0.420	-0.038
β_2	0.332	0.307	0.397	-0.358	1.252
β_3	0.321	0.304	0.391	-0.459	1.128

The Bayesian point estimate of the coefficient reflecting the effect of *Actual Lead* exposure on *Ciq* is negative, and since the central 95% region of the posterior lies between -0.420 and -0.038, I conclude that exposure to environmental lead is indeed deleterious according to this model and my prior uncertainty over the parameters.

[8] For details about the Gibbs sampler implementation, see Scheines, et al., 1998.

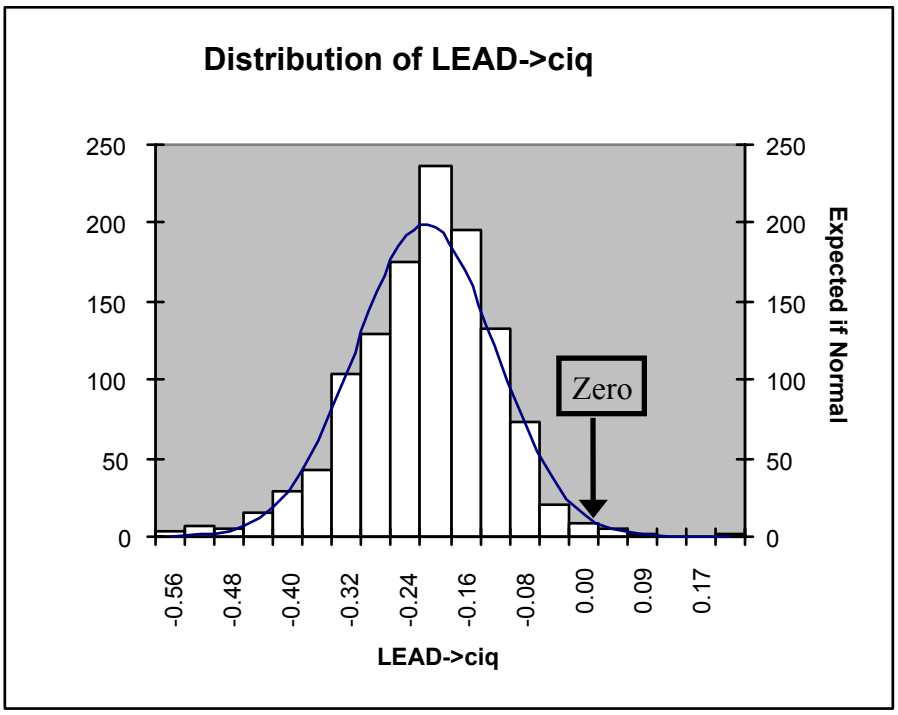

Figure 3. Histogram of relative frequency of β_1 in Gibbs sample

Although my uncertainty about the amount of measurement error associated with *Med* and *Piq*, which are proxies for environmental stimulation and genetic factors respectively, is not sufficient to make β_1 insignificant, it is sufficient to make β_2 and β_3 insignificant. That is, the central 95% of the sample from the posterior over both β_2 and β_3 includes 0. Since these coefficients represent the effect of environmental stimulation and genetic factors on a child's cognitive abilities, it seems reasonable to insist that they are at least positive in sign. I thus reran the analysis, but imposed 0 as a lower bound on β_2 and β_3. The posterior distribution over β_1 was slightly less diffuse, and centered over roughly the same value.

In fact I sampled from several posteriors corresponding to different priors, and in each case I got similar results. Although the size of the Bayesian point estimate for *Acutal Lead*'s influence on *Ciq* moved up and down slightly, its sign and significance (the 95% central region in the posterior over β_1 was always below zero) were robust.

I also ran the Gibbs sampler on an errors-in-all-variables model that included all six of Needleman's original regressors. In this case the bounds Klepper derived proved

important. Recall that the measurement error on *Fab* was required to be below .06. Using a prior in which substantial mass violated this bound, the sampler did not converge.

Table 4. Informative part of the prior in the errors-in-all-variables model including all six original regressors.

Parameter	Mean (μ_0)	Standard Deviation (σ_0)
$\sigma^2(\varepsilon_{Lead})$	0.05	0.05
$\sigma^2(\varepsilon_{Med})$	0.10	0.10
$\sigma^2(\varepsilon_{Piq})$	0.10	0.10
$\sigma^2(\varepsilon_{Fab})$	0.05	0.05
$\sigma^2(\varepsilon_{Mab})$	0.05	0.05
$\sigma^2(\varepsilon_{Nlb})$	0.05	0.05

Using a prior that was uninformative except for the parameters I show in Table 4, the histogram of values for β_1 in the Gibbs sample (Figure 4) was substantially different than the one in Figure 3.

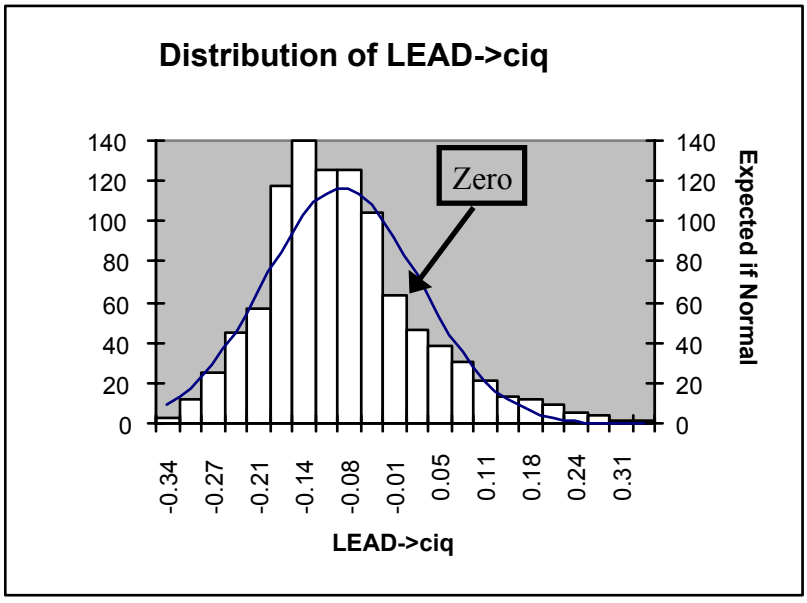

Figure 4. Gibbs sample from model with six regressors.

A full Bayesian analysis would incorporate uncertainty over these and other model specifications, and in future work I intend to address this problem. Given the two errors-in-all-variables models I have considered here, however, I am highly inclined to favor the smaller model suggested by TETRAD II's analysis. Given this model, which is perfectly plausible, the data quite clearly support Needleman's original conclusion.

References

Casella, G., & George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician,* 46*,* 167-174.

Grant, L., et al., (1983). "Draft air lead criteria document" (Environmental Protection Agency, Washington, D.C., 14 November, appendix 12-c.

Glymour, C., Spirtes, P. and Scheines, R. (1994). "In Place of Regression," in *Patrick Suppes: Scientific Philosopher*, Paul Humphreys (editor), Vol. 1, Kluwer Academic Publishers, Dordrecht, Holland.

Kiiveri, H. and Speed, T. (1982). Structural analysis of multivariate data: A review. *Sociological Methodology*, Leinhardt, S. (ed.). Jossey-Bass, San Francisco.

Klepper, S. (1988). Regressor diagnostics for the classical errors-in-variables model. *Journal of Econometrics*, 37, 225-250.

Klepper, S., & Leamer, E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica*, 52, 163-183.

Klepper, S., Kamlet, M., and Frank, R. (1993) Regressor Diagnostics for the Errors-in-Variables Model - An Application to the Health Effects of Polution, *Journal of Environmental Economics and Management*. 24, 190-211.

Needleman, H., et. al, (1979). *New England Journal of Medecine*, 300, 389.

Needleman, H., Geiger, S., and Frank, R. (1985). "Lead and IQ Scores: A Reanalysis," *Science*, 227, pp. 701-704.

Robins, J., and Wasserman, L. (1996). On the Impossibility of Inferring Causation from Association Without Background Knowledge, *Unpublished manuscript*, CMU Dept. of Statistics, Pittsburgh, PA.

Scheines, R. (1993). Causation, Indistinguishability, and Regression. *Softstat '93: Advances in Statistical Software 4*. pp. 89-99. Gustav Fischer, New York.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). "Bayesian Estimation and Testing of Structural Equation Models," *Psychometrika* 64, 1, pp. 37-52.

Scheines, R., Spirtes, P., Glymour, G., & Meek, C. (1994). *TETRAD II: Tools for causal modeling. User's manual*. Hillsdale, NJ: Erlbaum.

Smith, A.F.M., & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B,* 55, 3-23

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer.