Estimating Latent Causal Influences: TETRAD II Model Selection and Bayesian Parameter Estimation

Richard Scheines
Department of Philosophy, Carnegie Mellon University

Abstract

The statistical evidence for the detrimental effect of low level lead exposure 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 proved crucial in making the statistical evidence for the accepted epidemiological conclusion seem decisive. The first is a variable-selection routine in TETRAD II, 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 statistical causal inference in which two pieces of artificial intelligence technology proved crucial. The pieces are TETRAD II's Build module applied to a variable selection problem in linear regression, and TETRAD III's Gibbs sampler[1] applied to approximating the posterior distribution over the parameters of an "underidentified" linear model of the effect of lead exposure on IQ.

By measuring the concentration of lead in a child's baby teeth, Herbert Needleman was the first epidemiologist to even approximate a reliable measure of cumulative lead exposure. His work helped convince the United States to eliminate lead from gasoline and most paint (Needleman, et. al., 1979). Needleman's original statistical analysis of data he and colleagues collected on lead exposure and IQ scores (basically ANOVA) was criticized by the EPA (Grant, et al., 1983), which concluded that his data neither supported nor rejected the conclusion that lead was toxic at the doses he recorded in asymptomatic children. Needleman reanalyzed his data with multiple regression, and found that even after controlling for five covariates, the estimated effect of lead on IQ was negative and significant (Needleman, et al., 1985).

This quieted the EPA, but aroused more sophisticated criticism 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

[1]TETRAD III is now under development. See the TETRAD Project Home Page for details: http://hss.cmu.edu/philosophy/TETRAD/tetrad.html

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 Needleman's 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 each measured regressor (Klepper, 1988, 1993). The required measurement error bounds vary with each problem, however, and those required in order to bound the effect of actual lead exposure below 0 in Needleman's model seemed wholly unreasonable. Klepper concluded that the statistical evidence for Needleman's hypothesis was indeed weak.

Reanalyzing Needleman's data and regression model, I used TETRAD II to eliminate three spurious covariates that Needleman's backwards step-wise procedure had erroneously included. In fact the variables that TETRAD II eliminated were precisely those which required unreasonable measurement error assumptions. 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 coefficients 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, 1995) to do a Bayesian estimation of the 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.

2. Variable Selection with TETRAD II

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 elimination regression, arriving at a final regression equation involving lead and five covariates. The covariates were measures of genetic contributions to the child's IQ (the parent's IQ), the amount of environmental stimulation in the child's early environment (the mother's education), physical factors that might compromise the child's cognitive endowment (the number of previous live births), and the parent's age at the birth of the child, which might be a proxy for many factors. The measured variables they used are as follows, with the correlations among these variables and the significance of each correlation given in Table 1.

ciq - child's verbal IQ score
piq - parent's IQ scores
lead - measured concentration in baby teeth
mab - mother's age at birth
med - mother's level of education in years
fab - father's age at birth
nlb - number of live births previous to the sampled child

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[2] 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 $R^2 = .271$.

$$\hat{ciq} = -.143\,\text{lead} + .219\,\text{med} + .247\,\text{piq} + .237\,\text{mab} - .204\,\text{fab} - .159\,\text{nlb} \qquad [1]$$
$$\qquad\qquad (2.32) \quad (3.08) \quad (3.87) \quad (1.97) \quad (1.79) \quad (2.30)$$

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: a sizable unconditional association between X and Y might not be due to a direct causal link from X to Y, but rather at least partly from confounders (common causes of X and Y), or intermediate causes; statistically controlling for covariates leaves only the true causal association between X and Y. 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.

[2] The covariance data for my reanalysis was originally obtained from Needleman by Steve Klepper, who generously forwarded it to me. In this, and all subsequent analyses, I use the correlation matrix.

Linearity is not the issue, rather it is whether a significant association between X and Y, controlling for *all* the other potential confounders, is the right test for a direct causal connection between X and 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 **X** is specified. Assuming that **X** is prior to Y, in which case Y cannot cause any $X \in \mathbf{X}$, we say that X is causally adjacent to Y relative to the set **X** just in case either X is a direct cause of Y relative to **X**, or there is a Z not in **X** such that Z is a common cause of X and Y. Two assumptions are all that is needed to settle the issue of whether any $X_i \in \mathbf{X}$ is causally adjacent to Y relative to **X** from population data: the Causal Markov condition and Faithfulness.[3] 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 assumption 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 just in case X and Y are dependent conditional on *every subset* of **X** - {X,Y} (Spirtes, et al., 1993). Contrast this criterion with the one used in multivariate regression: X is a direct cause of Y just in case X and Y are dependent conditional on *exactly the set* **X** - $\{X_i, Y\}$. The model in Figure 1, in which $\mathbf{X} = \{X_1, X_2, X_3\}$ and Z is unmeasured, makes the error in the regression criterion vivid.

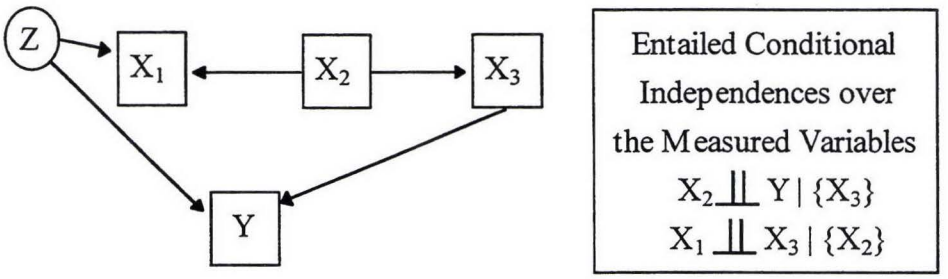

Figure 1: A model that fools regression

[3]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).

The 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 it 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.

It turns out that the regression criterion is reliable only when X is known to be prior to Y *and* the measured variables are known to be **confounder complete**,[4] i.e., all common causes of two variables in **X** U {Y} are already in **X** U {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 II (Scheines, et al., 1994) does not assume confounder completeness, and uses the correct criterion for causal adjacency under the Causal Markov and Faithfulness assumptions. It makes statistical decisions about independence by formulating null hypotheses, e.g., that $\rho_{X2,Y.X3} = 0$, and accepting them if they cannot be rejected at a user set significance level. In the lead example, the variables are distributed approximately multivariate normal, so I use partial correlations to test for conditional independence. In Figure 2 below I show the relevant part of the output from the FCI algorithm on the lead data, with a significance level of .05, and temporal information such that ciq is not prior to lead, and lead is not prior to every other regressor. The output indicates that only lead, med, and piq are adjacent to ciq.

```
List of vanishing (partial) correlations that made
TETRAD remove adjacencies.

Corr. :  Sample (Partial) Correlation
Prob. : Probability that the absolute value of the sample
(partial) correlation exceeds the observed value, on the
assumption of zero (partial) correlation in the population,
assuming a multinormal distribution.

Edge              (Partial)
Removed           Correlation              Corr.      Prob.
-------           -----------              -----      -----
fab -- ciq        rho(fab ciq)             -0.0003    0.9920
.
mab -- ciq        rho(mab ciq)              0.0540    0.4252
nlb -- ciq        rho(nlb ciq . med)       -0.1141    0.0914
```

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

```
-------------------------------------------------------
NOT assuming causal sufficiency
The Partial Ancestral Graph (PAG):
Significance Level =  0.0500

lead --> ciq
.
med o-> ciq
piq o-> ciq
```

Figure 2. Output from TETRAD II's Build Module on Needleman's data.

TETRAD II found that mab, fab, and nlb are *not* causally adjacent to ciq, contrary to Needleman's regression. In Needleman's data, these covariates are more highly 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, 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. Since the correct criterion for determining causal adjacency eliminates an adjacency between X and Y if they are independent conditional on *any* subset (including the empty set), TETRAD II eliminated the fab-ciq and mab-ciq adjacency because it accepted unconditional independence, and the nlb-ciq adjacency because it accepted the correlation between them as vanishing conditional on med. Asserting that the latent variable that nlb is a proxy for (e.g, Mother's Physical Factors) is not causally adjacent to ciq is a little more delicate. We must assume there is a connection between ciq and nlb through med, which seems implausible, or that med is highly correlated with the latent it proxies for (Environmental Stimulation), and that nlb and ciq are uncorrelated conditional on Environmental Stimulation, which is plausible.

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].

3. Estimating the Parameters of an "Underidentified" Model

As Klepper (1988, 1993) points out, and rightly so, 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 3 shows a full errors-in-all variables specification for the variables included by TETRAD II. The task is now to estimate the coefficient β_1.

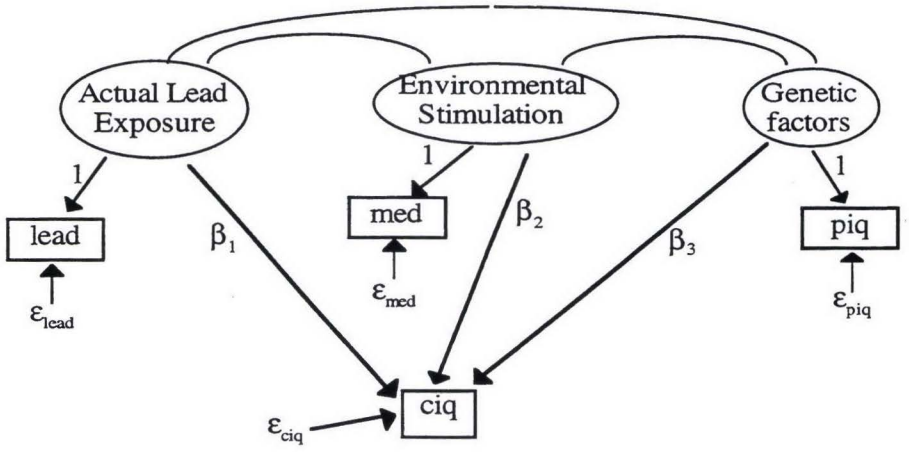

Figure 3. Errors-in-all-variables model for Lead's influence in IQ.

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 $\Sigma(\boldsymbol{\theta})$ that minimizes a discrepancy function of the implied and observed covariances, there are an infinity of parameterizations $\boldsymbol{\theta}'$ such that $\Sigma(\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta}')$. 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 Var(lead) = 1 = Var(Actual Lead) + Var(ε_{lead}), the proportion of measured lead's variance due to measurement error is just Var(ε_{lead}), which is between 0 and 1. Similarly for the other regressors. Using a linear regression to estimate β 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 bound the actual coefficients, at least in sign, by putting upper bounds on the amount of measurement error. In 1988 and again in 1993 Klepper argued that the upper bounds required by his method to solve this problem (with all six regressors) 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 even remotely 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 his analysis on the reduced set of regressors, 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 conditions for bounding, not necessary, and 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 ($SD(\theta_{EAP})$), 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, 1995).[5] One can then use the sample from the posterior to estimate the posterior statistics from their sample counterparts, i.e., $\hat{\theta}_{EAP}$, $\hat{\theta}_{MDAP}$, $SD(\hat{\theta}_{EAP})$, $\hat{\theta}_{.025}$, 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.[6]

To apply the Bayesian solution to Needleman's 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. From consultations with critics, I guess that between 0% and 40% of the variance in Needleman's proxy is from measurement error, with 20% a conservative best guess. For the measures of environmental stimulation and genetic factors, I am less confident, so I guess 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 by specifying the multivariate normal prior over the model's 13 parameters given in Table 2, with no covariation between any of the model's parameters in the prior. Notice that the prior is only informative about the three error variances that parameterize the amount of measurement error in Needleman's original proxies. With a

---

[5] A Gibbs sampler for computing the posterior over the parameters of a structural equation model is now available in a beta-version of TETRAD III. See the TETRAD Project Home Page for details: http://hss.cmu.edu/philosophy/TETRAD/tetrad.html

[6] For details about the Gibbs sampler implementation, see Scheines, et al., 1995.

standard deviation of 4.0 around the other parameters, the prior is effectively uninformative everywhere else. The means for the non-error variances are set to the regression estimates from equation [2].

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

Parameter	Mean (μ_0)	Standard Deviation (σ_0)
Var(ε_{lead})	0.200	0.10
Var(ε_{med})	0.300	0.15
Var(ε_{piq})	0.300	0.15
Var(ε_{ciq})	0.757	4.00
Var(Actual Lead)	0.800	4.00
Var(Environ. Stim.)	0.700	4.00
Var(Genetic Factors)	0.700	4.00
β_1	-0.177	4.00
β_2	0.251	4.00
β_3	0.253	4.00
Cov(Act. Lead, Env. Stim)	-0.136	4.00
Cov(Act. Lead, Gen. Fac)	-0.058	4.00
Cov(Env. Stim, Gen. Fac)	0.527	4.00

Using this 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 4 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.

	$\hat{\theta}_{EAP}$	$\hat{\theta}_{MDAP}$	SD($\hat{\theta}_{EAP}$)	$\hat{\theta}_{.025}$	$\hat{\theta}_{.975}$
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 IQ 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.

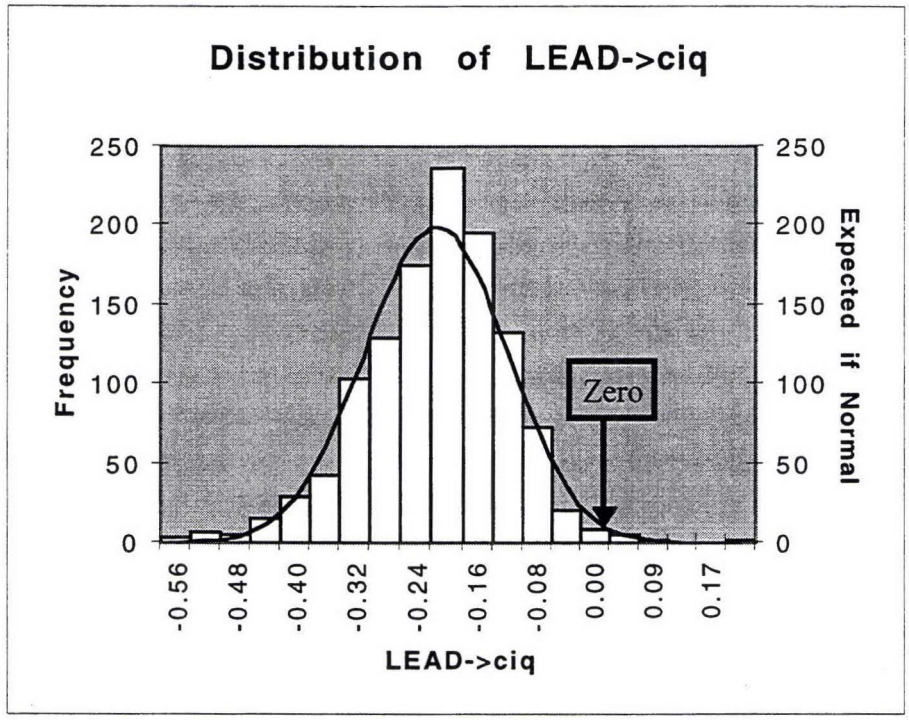

Figure 4. Histogram of relative frequency of β_1 in Gibbs sample

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 lead's influence on IQ 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)
$Var(\varepsilon_{lead})$	0.05	0.05
$Var(\varepsilon_{med})$	0.10	0.10
$Var(\varepsilon_{piq})$	0.10	0.10
$Var(\varepsilon_{fab})$	0.05	0.05
$Var(\varepsilon_{mab})$	0.05	0.05
$Var(\varepsilon_{nlb})$	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 5) was substantially different than the one in Figure 4.

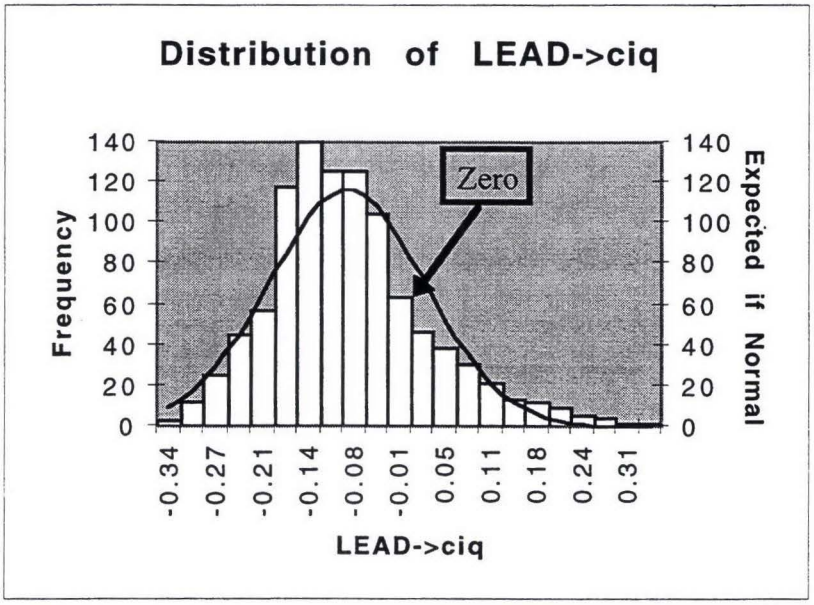

Figure 5. Gibbs sample from model with six regressors.

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 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. (1995). *Bayesian Estimation and Testing of Structural Equation Models*, (submitted to Psychometrika) *Technical Report CMU-PHIL-66*, Dept. of Philosophy, Carnegie Mellon Univ., Pgh, PA, 15213.

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.