THE DETERMINANTS OF READING SCORES: AN ANALYSIS BUILT ON EXPLICIT PRIOR INFORMATION*

By

Edward E. Leamer
Department of Economics
UCLA
Los Angeles, CA 90024

Discussion Paper Number 218
October 1981
Department of Economics
University of California, Los Angeles

*Prepared for a conference on methodology sponsored by the System Development Corporation. Research support from SDC and from NSF grant SOC 78-09479 is gratefully acknowledged.

THE DETERMINANTS OF READING SCORES: AN ANALYSIS BUILT ON EXPLICIT PRIOR INFORMATION

by
Edward E. Leamer

1.0 Introduction

Every analysis of associations in non experimental settings confronts a great excess of potential variables. The effectiveness of a data analysis depends fundamentally on how this morass is reduced to a manageable few. There are, I believe, three different approaches that may be taken. The first is to let the computer do the job. In the regression setting, this means using one of the variants of step-wise regression. Variables (or linear combinations in the case of principal component regression) are included if they "contribute adequately" to the explanation, and are excluded otherwise. The second approach is to let the researcher select the possible subsets of the variables for inclusion, to estimate these alternative "models" and to select the one or ones that work "best." The third approach begins with a formalization of a priori opinion about the size of the coefficients in the general model. The prior opinions are packaged in a probability distribution and the data are used via Bayes rule to form a posterior distribution.

Each of these approaches has serious shortcomings. The computer searches and the researcher searches are quite common in practice, but the Bayesian approach has remained almostly completely a theoretical toy (pipe dream?). It is my intention in this paper to demonstrate that the shortcomings of the Bayesian approach can be overcome at reasonable cost. I will try to convince you that the Bayesian approach, with certain important modifications, is actually the best of the three approaches, best by a considerable margin, I believe. In order to do this, I provide not just the recipe, but a taste of the pudding as well. That is, I will contrast a Bayesian analysis of reading scores with the two alternative approaches.

2.0 Computer Selection, Ad Hoc Selection or Bayes Selection?

The computer searches are obviously "mechanical" but deserve as well the epithet "uninterpretable." Step-wise regression, as it is usually practiced, has no foundation in statistical theory. The estimates and the standard errors which result have no meaning, or at least a meaning that has eluded theoretical statisticians. One problem is obvious with step-wise regression. If two variables are highly correlated, one will be excluded, and one included, the included playing the role of itself plus the role of the excluded variable. Suppose you wanted to explain school-average reading percentiles in terms of the two variables: x_1, the ratio of the number of black teachers to the number of pupils and, x_2, the ratio of white teachers to the number of pupils. Because these variables are highly correlated, step-wise regression will select one for inclusion and the other for exclusion. The model that results will make little sense. If the black teacher variable is included, the model will over-estimate the effect of black teachers and underestimate the effect of white teachers.

Economists, to their credit, do not often use step-wise regression. Instead, they try different subsets of the variables and select the model that "looks best." To continue the example, they might first regress reading percentile on the total teacher/pupil ratio (x_1+x_2), and then try including both variables separately, x_1 and x_2. If the second equation produces different signs for the two coefficients, if the coefficients are not "significant," and if the R^2 is not much better than the first equation, then they may revert to the first.

The step-wise procedure produces the non-sensical result

$$P = \hat{\beta}_o + \hat{\beta}_1 x_1 \qquad (1)$$

The researcher-selection procedure produces one of two sensible outcomes

$$P = \hat{\beta}_0 + \hat{\beta}(x_1+x_2) \quad , \text{ or} \tag{2}$$

$$P = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad , \tag{3}$$

The second equation is selected if $\hat{\beta}_1$ and $\hat{\beta}_2$ are "close" and if the hypothesis $\hat{\beta}_1 = \hat{\beta}_2$ can be rejected at traditional levels. If $\hat{\beta}_1 = \hat{\beta}_2$ can be rejected, and if $\hat{\beta}_1$ and $\hat{\beta}_2$ are not close, say opposite in sign, then you select another variable, say x_3, which is sufficiently correlated with either x_1 or x_2, that the regression

$$P = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \tag{4}$$

is "OK."

By now you should realize what is wrong with the researcher-selection methods. They are ad hoc. They are whimsical. For reasonably complicated problems they cannot be fully reported. They depend on the researcher's judgment, in this case the opinion that β_1 and β_2 are close. The final model is therefore an unknown mixture of the data information and the implicit prior information. There really is no way to tell from the reported equations how many were tried and what the researcher's prior was. There is, furthermore, no way for the researcher to know if he is using his prior information correctly.

The third method, the Bayesian approach, seems ideally suited to this problem. The data alone probably cannot be used to estimate sensibly β_1, β_2 and β_3. We can augment the data information with explicit prior

information. It is computationally most convenient if the prior for β_0, β_1, β_2, and β_3 is in the form of a multivariate normal distribution. To use the same information as the ad hoc searcher, the difference $\beta_1 - \beta_2$ can be taken to have mean zero and variance V_1, and β_3 can be taken to have mean zero and variance V_2. All other linear combinations, except those of the form $a(\beta_1-\beta_2) + b\beta_3$, can be taken to have infinite variance, to reflect a state of relative ignorance about other parameters. The problem with the Bayesian approach now arises. What numerical values should be selected for V_1 and V_2? No one that I know can select these numbers comfortably. Under the assumption of normality, the interval $-1.96V_1^{\frac{1}{2}} < \beta_1 - \beta_2 < 1.96V_1^{\frac{1}{2}}$ has probability .95. In order to select V_1, you must ask yourself how different β_1 and β_2 can be. These coefficients are changes of the reading score in response to changes in the teacher-pupil ratio. As for myself, when I try to select V_1 I start to think about $(\beta_1+\beta_2)/2$. It is hard to think about $\beta_1 - \beta_2$ without thinking about the absolute size of β_1 and β_2. To be explicit, let us write the model as

$$P = \beta_0 + \left(\frac{\beta_1+\beta_2}{2}\right)(x_1+x_2) + \frac{\beta_1-\beta_2}{2}(x_1-x_2) + \beta_3 x_3 \,.$$

$$\equiv \theta_0 + \theta_1(x_1+x_2) + \theta_2(x_1-x_2) + \theta_3 x_3 \,.$$

Suppose that you decide $(\beta_1+\beta_2)/2$ has mean M_3 and variance V_3; by that I mean you are willing to announce that the interval $M_3 - 1.96V_3^{\frac{1}{2}} < (\beta_1+\beta_2)/2 < M_3 + 1.96V_3^{\frac{1}{2}}$ has prior probability .95. It may then make sense to think that β_1 and β_2 are different by no more than 10 per cent of the mean: $|\beta_1 - \beta_2| < .1(|\beta_1|+|\beta_2|)/2$. Possibly you could represent this by selecting $1.96V_1^{\frac{1}{2}} = .1M_3$; or maybe to allow for your prior uncertainty about $(\beta_1+\beta_2)/2$ you should let $1.96V_1^{\frac{1}{2}} = .1(M_3+1.96V_3^{\frac{1}{2}})$.

Next we have to select M_3 and V_3, the mean and variance for $\theta_1 = (\beta_1+\beta_2)/2$. If S is the percentile reading score, then θ is the improvement in the percentile induced by changing the teacher-pupil ratio by one. An experiment that would have more meaning would be to increase the teacher-pupil ratio from 1/50 to 1/25. The variable $x_1 + x_2$ increases by $.02 = 1/25 - 1/50$, and S increases by $.02\theta_1 = .02(\beta_1+\beta_2)/2$. Now ask yourself by how much would the reading scores improve if the teacher-pupil ratio increased from 1/25 to 1/50. That seems to me to be a big improvement in the quality of education and I would imagine a big increase in reading scores. A big increase in scores on a year-to-year basis is, I suppose, 5 percentiles. But something is wrong here; if you start at the 98th percentile, you can hardly improve by 5 percentiles. And if you improved the teacher-pupil ratio and maintained it, you can hardly improve by 5 percentiles indefinitely. A change in the functional form is required, and one that seems to make sense is

$$100P/(P-100) = \theta_0 + \theta_1(x_1+x_2) + \theta_2(x_1-x_2) + \theta_3 x_3 \ . \qquad (5)$$

The transformation $S = 100P/(100-P)$ stretches the interval from 0 to 100 into the interval from zero to infinity. The derivative of S with respect to P is $(100)^2/(100-P)^2$. This is one at $P = 0$ and increases monotonically as P approaches 100. This means that a given change in the teacher-pupil ratio will have the largest effect at low reading percentiles and will have an increasingly imperceptible effect as P approaches 100. An alternative dependent variable is ln S where ln is the natural logarithm. Table 1 compares the changes in P associated with equal changes in S with the changes in P associated with equal change in ln S. I believe that the former

steps in P better represent the outcome of equal amounts of teaching effort, and I will adopt the first dependent variable, although for statistical reasons ln S could be better.

Next we have to decide if the linear function $x_1 + x_2$ is better than some non-linear function, such as $\ln(x_1+x_2)$. If T represents the number of teachers and P the number of pupils, then the linear form $S = \beta_0 + \beta_1 T/P$ has derivative $\partial S/\partial T = \beta_1/P$. This means that the marginal effectiveness of an extra teacher decreases with the number of pupils but is independent of the number of teachers. On the other hand, the log-linear form $S = \beta_0 + \beta_1 \ln(T/P)$ has derivative $\partial S/\partial T = \beta_1/T$, which decreases with the number of teachers. The log-linear form thus includes a saturation effect as the number of teachers increases. It seems to me to be a more sensible characterization of the school learning environment. In order to allow for differences in the effectiveness of black and white teachers, you could use the functional form $S = \beta_0 + \beta_1 \ln(x_1+\beta_2 x_2) + \beta_3 x_3$. This form is non-linear in the parameters and is consequently difficult to estimate. A less desirable functional form that is linear in parameters is

$$S = 100P/(100-P) = \gamma_0 + (\gamma_1+\gamma_2 \frac{x_1}{x_1+x_2})\ln(x_1+x_2) + \gamma_3 x_3 \qquad (6)$$

Having iterated to this new, more sensible functional form, let us now think about the prior distribution for γ_1 and γ_2. The parameter γ_1 is the change of S with respect to a percentage change in the teacher-pupil ratio if all teachers are white. A large change in S is 50 units. That corresponds to a change in the reading percentile from 0 to 33, or 33 to 50, or 50 to 60. To get such a dramatic change in the reading score, you would have to double the number of teachers (I guess). This allows me to solve for a guesstimate of γ_1 : $50 = \gamma_1(\ln 2x - \ln x)$, or $\gamma_1 = 50/\ln 2 = 72$. This will

serve as my prior mean. I am not an educational expert and I am very uncertain about γ_1. Accordingly, I will select the prior variance of γ_1 to be $(25)^2$, thereby asserting that γ_1 is quite likely (probability .95) to be in the interval $72 - 1.96 \cdot 25 \leq \gamma_1 \leq 72 + 1.96 \cdot 25$. Because I don't think there is a substantial difference between black and white teachers I will select the prior mean of γ_2 to be zero and the prior variance to be $(5)^2$, thereby assigning prior probability .95 to the interval $-5 \cdot 1.96 < \gamma_2 < 5 \cdot 1.96$. Similar considerations would produce a prior for γ_3.

If you are tired from all this, a bit confused, and rather incredulous, then I have made important point number one. If you think the functional form (6) is likely to be a better representation of the data than (4) and if you think you know something about the parameters in (6) but have a hard time thinking about the parameters in (4), then I have made important point number two.

Point number two is that there is a substantial side benefit to a Bayesian approach. In order to think about prior distributions for the parameters of a model, the parameters have to make sense. You are forced into choosing a functional relationship that can be understood. You will identify parameter values that are extreme, and you will be in a position to be surprised by the data. (It is surprises that allow creative learning.)

Point number one is that it is difficult, if not impossible, to be a textbook Bayesian. There has to be a reason why a formal Bayesian analysis is not used in practice, when judgment is so often used and so often reported as the reason for selecting one set of estimates over another. The reason, I believe, is that researchers are unable to form complete prior

distributions. Even for the relatively simple problem reported above there were so many twists and turns, so many partly arbitrary decisions, that you have to feel uncomfortable with the final product. You ought to be uncomfortable as well with the ad hoc selection procedure, but most people are blissfully ignorant, and act as if the equation they select was really the only one that matters anyway.

The obvious benefit of a Bayesian approach is that you make overt use of fully articulated prior information. You understand exactly what you have done, and so does your reader, at least to the extent that a prior probability distribution is understandable. But that's the rub. It's difficult to select the prior distributions, and it is logically impossible to understand fully any given distribution. The cumulative distribution for a single parameter β is $F(c) = Pr(\beta \leq c)$. In principle, you have to select $F(c)$ for every value of c. That will take an unlimited amount of time. Instead, the cumulative is parameterized in terms of a few parameters, say normal with mean M and the variance V. These parameters are then selected by introspection concerning a couple of points of the cumulative distribution, say $.025 = F(c_1)$ and $.975 = F(c_2)$, which imply $M + 1.96V^{\frac{1}{2}} = c_1$ and $M - 1.96V^{\frac{1}{2}} = c_2$. In so doing, you have selected the complete cumulative distribution, even though virtually every point has not been given the slightest thought. Moreover, even the values of c_1 and c_2 are difficult to choose. I can't, and I doubt that you can, distinguish between probabilities .01, .02 and .03. It could well be that the value of c_1 that I select in fact implies $.01 = F(c_1)$ rather than $.025 = F(c_1)$. I really don't have any way of knowing.

If you are now prepared to discard forever the Bayesian approach, please recall that I promised to demonstrate that the shortcomings can be

overcome. The way to do this is a sensitivity analysis. There are two kinds of sensitivity analyses. A local sensitivity analysis studies the changes in the posterior distribution induced by small changes in the prior distribution. A global sensitivity analysis studies the changes in the posterior distribution induced by large changes in the prior. In each case it is up to the researcher to identify interesting perturbations, large or small.

Let us take the prior for the vector of coefficients β to be multivariate normal with mean vector m and variance-covariance matrix V. Given the usual normal linear regression model, it is then possible to compute the posterior distribution $f(\beta|Y,m,V)$ where Y stands for the observed data. Various features of the posterior distribution, such as the mean $E(\beta|Y,m,V)$, the mode $M(\beta|Y,m,V)$, or the probability $Pr(\beta_i \geq 0|Y,m,V)$ are natural candidates for reporting. In a local sensitivity analysis you would report the derivatives of these quantities with respect to the prior mean vector m and the prior variance-covariance matrix V.

A global sensitivity analysis proceeds somewhat differently. We begin with a rather general family of prior distributions, say all those with a given mean vector m. Corresponding to this family of priors is a family of posteriors. Extreme members of this class are identified and bounds for the issues of interest are reported, say

$$\min_V E(\beta_i|Y,m,V) \leq E(\beta_i|Y,m,V) \leq \max_V E(\beta_i|Y,m,V), \text{ or}$$

$$\min_V Pr(\beta_i \geq 0|Y,m,V) \leq Pr(\beta_i \geq 0|Y,m,V) \leq \max_V Pr(\beta_i \geq 0|Y,m,V) .$$

If these bounds are too wide to be useful, you must seek ways credibly to narrow the family of priors. One possibility is to restrict the variance

matrix from above and below, $V_* \leq V \leq V^*$, where $A \leq B$ means $B - A$ is positive definite. The narrower is the "interval" between V_* and V^*, the smaller will be the family of posterior distributions and the tighter will be the bounds for the issues of interest. The ideal outcome is that you can identify a family of priors so broad that with virtual certainty your prior is a member, but so narrow that the specification intervals for the issues of interest are useful. If these intervals are wide, and the family of priors cannot be credibly reduced, then inference with your data set is suspended.

The point of a global sensitivity analysis is to base inference only on assumptions you can comfortably make. I am not comfortable with the assumption that the prior for β is a multivariate normal distribution with mean vector m and covariance matrix V. There are settings in which I am comfortable saying the prior is located at m. I also think it is very informative to analyze the data by bounding the prior covariance matrix from above and below, $V_* < V < V^*$, but I am somewhat unhappy with the assumption of normality which underlies this bound. (I am equally uncomfortable with the assumption of normality for the error distribution (sampling process).) I often have feelings about signs of coefficients and I would like to be able to use prior bounds of the form $P_* \leq \Pr(\beta_i > 0) \leq P^*$. Computer codes that make use of the prior bound $P_* < \Pr(\beta_i \geq 0) \leq P^*$ or that compute posterior bounds for $\Pr(\beta_i \geq 0 | Y)$ are not yet available, and in this paper I will present posterior bounds for $E(\beta_i | Y, m, V)$ based on prior bounds for V.

2.0 Mathematical Results

The mathematical results that are the foundation of the analysis to be presented are based on the assumptions that: (1) the observable $T \times 1$ vector Y has a multivariate normal distribution with mean $X\beta$ and variance

$\sigma^2 I$, where X is a $T \times k$ observable matrix, β is a $k \times 1$ vector of unobservable parameters, and σ^2 is a scalar variance parameter; (2) the vector β has a multivariate prior distribution which is normal with mean m and variance V. From this joint distribution of Y and β it is straightforward to compute the conditional distribution of β given Y, which is normal with moments

$$\hat{\beta}(V,m) \equiv E(\beta|Y,V,m,\sigma^2) = (H+V^{-1})^{-1}(Hb+V^{-1}m) \qquad (7)$$
$$Var(\hat{\beta}|Y,V,m,\sigma^2) = (H+V^{-1})^{-1}$$

where

$$H = \sigma^{-2}X'X \ ,$$
$$Hb = \sigma^{-2}X'Y \ .$$

Although these moments condition on σ^2, an unknown parameter, I will be acting as though $s^2 = (Y-Xb)'(Y-Xb)/(T-k)$ were a perfect estimate of σ^2, and for computational purposes replace σ^2 by s^2 in every formula. Formally speaking, I will be acting as if the product of a normal kernel and a Student kernel can be well approximated by the product of two normals. This requires that the degrees of freedom T - k be large enough that σ^2 is accurately estimated.

The local sensitivity analysis discussed in Section 3 proceeds straightforwardly by differentiating (7) with respect to the prior mean m:

$$\partial\hat{\beta}/\partial m = (H+V^{-1})^{-1}V^{-1} \ .$$

This matrix of derivatives can be collapsed into a scalar when interest focuses on some single function of β, say $\psi'\beta$ where ψ is a vector of constants. Then

$$\partial\psi'\hat{\beta}/\partial m = \psi'(H+V^{-1})^{-1}V^{-1} \quad . \tag{8}$$

The global sensitivity analysis makes use of the following result from Leamer (1981). In these results the prior mean is taken to be the vector zero without loss of generality.

Theorem 1: Given that $V \geq V_*$ with V_* positive definite, then $\hat{\beta}(V)$ lies in the ellipsoid

$$(\hat{\beta} - f_*)'H_*(\hat{\beta} - f_*) \leq c_* \tag{9}$$

where

$$\begin{aligned} H_* &= HV_*H + H \\ f_* &= (HV_*H + H)^{-1}(HV_*Hb + Hb/2) \\ &= (H + V_*^{-1})^{-1}(Hb + V_*^{-1}b/2) \\ c_* &= b'V_*^{-1}(H + V_*^{-1})^{-1}Hb/4. \end{aligned}$$

Conversely, any point in this ellipsoid is a posterior mean $\hat{\beta}(V)$ for some $V \geq V_*$.

Theorem 2: Given that $V \leq V^*$ with V^* positive definite, then $\hat{\beta}(V)$ lies in the ellipsoid

$$(\hat{\beta} - f^*)'H^*(\hat{\beta} - f^*) \leq c^* \tag{10}$$

where

$$\begin{aligned} H^* &= (V^{*-1} + H) \\ f^* &= (V^{*-1} + H)^{-1}Hb/2 \\ c^* &= b'H(V^{*-1} + H)^{-1}Hb/4. \end{aligned}$$

Conversely, any point in this ellipsoid is a posterior mean $\hat{\beta}(V)$ for some $V \leq V^*$.

Theorem 3: Given that $V_* \leq V \leq V^*$ with V_* and V^* positive definite, then $\hat{\beta}(V)$ lies in the ellipsoid

$$(\hat{\beta}-\hat{f})'\hat{H}(\hat{\beta}-\hat{f}) \leq \hat{c} \qquad (11)$$

where

$$\hat{H} = (H+V^{*-1})(V_*^{-1}-V^{*-1})^{-1}(H+V^{*-1}) + (H+V^{*-1})$$

$$\hat{f} = [(H+V^{*-1}) + (V_*^{-1}-V^{*-1})]^{-1}(Hb+(V_*^{-1}-V^{*-1})(H+V^{*-1})^{-1}Hb/2)$$

$$\hat{c} = b'H(H+V^{*-1})^{-1}(V_*^{-1}-V^{*-1})(H+V^{*-1}+V_*^{-1}-V^{*-1})^{-1}Hb/4$$

Ellipsoids (F), (9), (10) and (11) are depicted in Figure 1. Ellipsoid F allows V to be any matrix and has a boundary that is generated by priors which assign zero prior variance to some linear combination of parameters but are otherwise diffuse. In other words, these are least-squares estimates subject to linear restrictions. These boundary points are not obtainable if the prior variance matrix is bounded from above or from below, with the exception of the origin in the former case and least-squares in the latter.

As sample size grows, the outside ellipsoid (F) does not collapse, the ellipsoid (10) which contains the origin grows to fill all of (F), and the ellipsoid (9) which contains the least-squares point collapses to that point. Thus, in order for large samples to make priors irrelevant it is necessary (of course) to exclude dogmatic priors, that is to bound the prior variance away from zero. Provided that dogmatic priors are excluded, $0 < V_*$, the ellipsoid (11) as sample size grows will eventually shrink to the least-squares point.

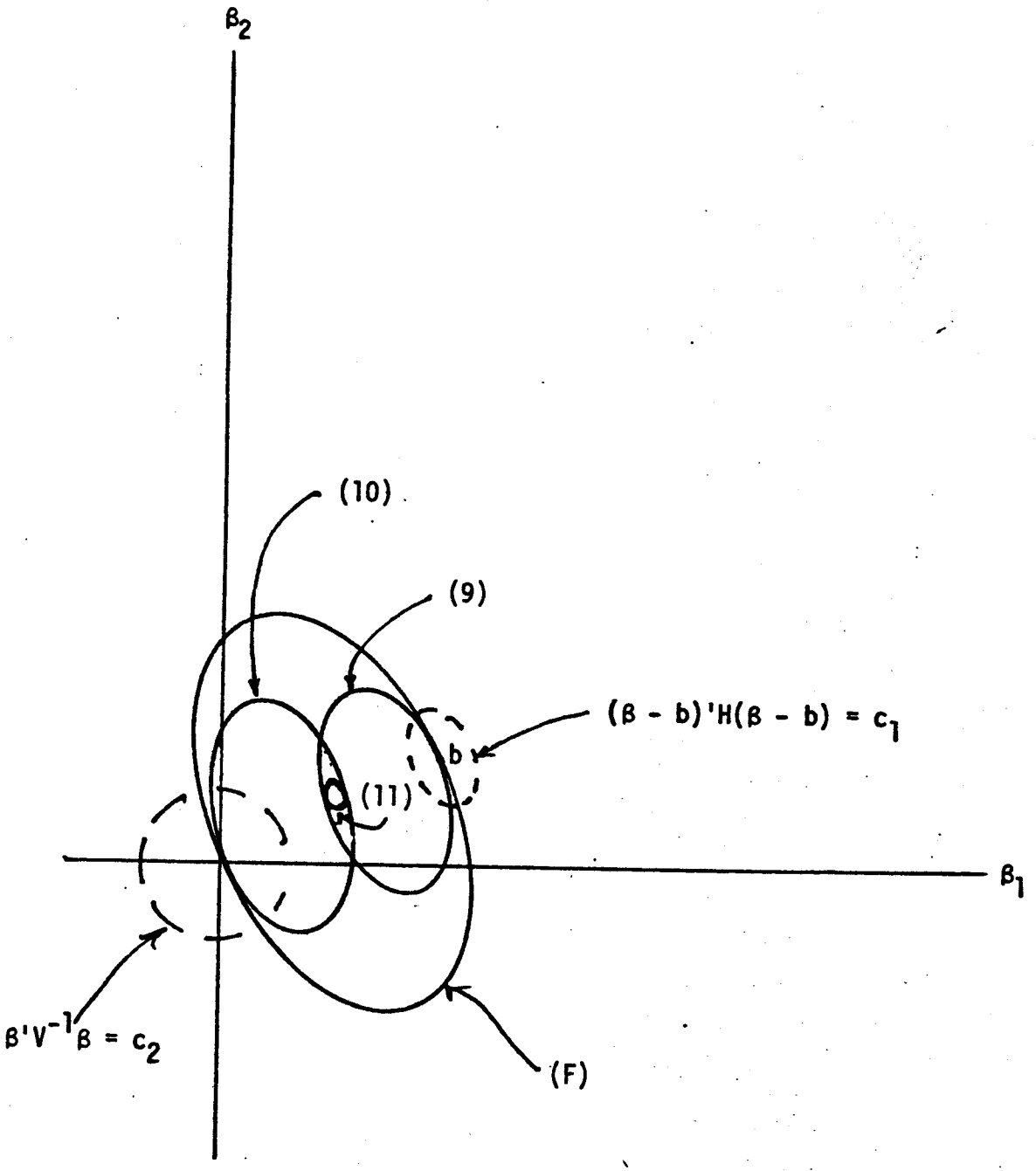

FIGURE 1: Ellipsoidal Bounds for Families of Prior Variance Matrices

(F) $0 < V$

(9) $V_* < V$

(10) $V < V^*$

(11) $V_* < V < V^*$

3.0 An Analysis of Reading Score Data

To illustrate the Bayesian techniques, I present an analysis of the determinants of elementary school reading scores in the Los Angeles Unified School District. The data were kindly provided by J. Ward Keesling. Included in the data set are reading scores for each of 436 elementary schools beginning in 1969 and ending in 1975. Also included are a fairly long set of variables which describe characteristics of schools such as race, and average family background.

The basic model which I will study takes the reading percentile in the spring of year t, denoted by P, to depend on the reading percentile the previous spring, P_0, the teacher/pupil ratio, x_1, and the non-instructional expenditure/pupil ratio, x_2:

$$100P/(100-P) = \delta 100P_0/(100-P_0) + \beta + \theta \ln x_1 + \gamma \ln x_2,$$

where δ, β, θ and γ are parameters to be estimated. The transformation $100P/(100-P)$ has the effect of stretching out the high percentiles. This reflects the difficulty of inducing schools with high scores to improve even more. In the absence of instruction, the current reading percentile can be expected to be somewhat lower than the previous percentile, largely because instructed students are improving their skills and "raising the curve" on which the percentiles are computed. The quality of instruction is measured in terms of the pupil/teacher ratio and the expenditure/pupil ratio. My suspicion is that reading achievement depends primarily on the size of classrooms and only incidentally on expenditures other than teaching salaries such as the hiring of administrators or the purchase

of equipment. Ideally, the term x_2 would include expenditures for classroom materials. Since it does not, γ must be expected to be a rather small number.

The parameters β, θ and γ are allowed to depend on seven variables: three racial composition variables, the logarithm of average family income, adult educational attainment, crimes, and pupil stability. These variables affect β because they measure to some extent the reading training that takes place outside of school. The variables affect θ and γ because they measure to some extent the quality of the school environment and the receptiveness of students to reading training. It is anticipated that some schools will be so affected by student turnover, criminal behavior, and adult indifference that no matter what quality of reading training is offered little learning can occur. In that case θ and γ would be very small numbers.

These variables are listed and more fully defined in Table 2. You will find there also a list of variables which were excluded from this study. The general research strategy which I am recommending begins with a model which is over-parameterized by traditional standards. The model which I have selected has twenty-five variables, including the constant and interactions. I do not expect that an ordinary least-squares regression with this model would make much sense. I do expect the Bayes estimates

to be reasonable, and the problem of "over-fitting" is cured by the use of a proper prior distribution. But you may now ask why stop at twenty-five variables? If you included the twenty other variables in Table 1 in the interactive form the model would include eight-four variables. Actually, I do not think that eighty-four is too many at a conceptual level. However, my computer program can handle only twenty five. By excluding these fifty-nine variables, I don't think I will be greatly influencing my inferences from this data set. Partly their effects are small and partly their effects are already part of the model. But there are many interesting hypotheses which I regrettably have chosen to ignore. For example, do black students respond better to black teachers than to white teachers? Are unhappy teachers ineffective? By listing these excluded variables, I hope I have made clear that any statistical analysis for reasons of costs operates within a horizon. We must accordingly reserve the right to extend that horizon when data peculiarities suggest.

The model including interaction terms is

$$S = \delta S_0 + (\beta_0 + \sum_{i=1}^{7} \beta_i z_i) + (\theta_0 + \sum_{i=1}^{7} \theta_i z_i) \ln x_1 + (\gamma_0 + \sum_{i=1}^{7} \gamma_i z_i) \ln x_2$$

where $S = 100P/(100-P)$ and $S_0 = 100P_0/(100-P_0)$. The issues of interest which I will discuss are the nine derivatives

$$dS/d \ln x_1 = \theta_0 + \Sigma \theta_i z_i$$

$$dS/d \ln x_2 = \gamma_0 + \Sigma \gamma_i z_i$$

$$dS/dz_i = \beta_i + \theta_i \ln x_1 + \gamma_i \ln x_2 \qquad i = 1, \ldots, 7.$$

The first two of these derivatives measures the effectiveness of teachers and expenditures in raising reading scores. These depend on the characteristics

of the schools. I will provide estimates of these parameters for the "average" school and for several extreme schools. The remaining seven derivatives will be evaluated only at the average values of $\ln x_1$ and $\ln x_2$.

In order to perform a Bayesian analysis it is necessary to form a prior distribution for the 25 parameters. This is done by selecting prior means and variances for 25 linear combinations of parameters, which are a priori distributed independently of each other. You regard a parameter θ_1 to be independent of θ_2 if, given information about θ_2, you do not change your mind about θ_1. The twenty five linear combinations of parameters which I take to be independent are listed in Table 3. There are sixteen of the individual parameters and 9 linear combinations indicated by dS/dz_i or dS/dx_i. These are derivatives of S evaluated at the data means of $\ln x_i$ and z_i.

It is appropriate, actually desirable, for you to question each aspect of my prior distribution. If this distribution greatly misrepresents your own opinions, then the sensitivity analysis that will be discussed will be less useful. I have not been engaged in research on reading achievement, and I ought not be expected to have the opinions of experts in the field. Another interesting approach would be to interrogate experts to elicit a more informed prior.

In Table 2, I have reported means and two times the standard errors for each linear combination of parameters. The linear combination is thought to be highly probable (.95) to be within two standard errors of its mean. Explanations for these choices are as follows:

(1) S_0: The coefficient of last year's reading score is quite likely to be in the interval from .8 to 1.2.

(2) β_0: One way to think about the constant is to suppose that z_i and $\ln x_i$ are all zero in which case the model becomes

$S = \delta S_0 + \beta_0$, but this is taking the functional form too seriously. It is better to think of β_0 as being selected to make the estimated function approximate in the region of the data as accurately as possible the true functions. This complicated task will ordinarily leave you with little knowledge of the constant. Hence the prior standard error is set to infinity.

(3) dS/dz_1: Referring to Table 1, a large change in S is 50 units. Steps of fifty units in S correspond to changes in the percentile from 33 to 50 to 60 to 66 to 71. As the racial composition varies from zero percent Spanish to 100 percent Spanish, I doubt that year-to-year changes in reading scores would vary by more than 100 units.

(4) dS/dz_2: Same as (3).

(5) dS/dz_3: Same as (3).

(6) dS/dz_4: Each 10% change in family income, could affect year-to-year changes in reading scores by no more than 5 units.

(7) dS/dz_5: Each grade change in adult education level is unlikely to affect year-to-year changes in reading scores by more than 10 units.

(8) dS/dz_6: A change in the crime rate per student from n per thousand to n + 1 per thousand is unlikely to affect year-to-year changes in reading scores by more than 5 units.

(9) dS/dz_7: When the percent of students who remain in the school varies by 1, year-to-year changes in reading scores are unlikely to vary by more than one unit.

(10) $dS/d\ln x_1$: A one percent increase in the teacher/pupil ratio is likely to increase year-to-year changes in reading scores by roughly 2.5. This change is unlikely to be outside the interval from -2.5 to 7.5.

(11) $dS/d\ln x_1$: The expenditure effect is expected to be half the teacher effect. A one-percent increase in expenditures is unlikely to change reading scores by more than 6.25 or less than -3.75.

(12) θ_i, γ_i, $i \neq 0$: All the interactive parameters have mean zero. My uncertainty about these parameters is similar to my uncertainty about the derivatives dS/dz_i.

Least-square estimates, three stepwise estimates and Bayes estimates of the model for grade 2 are reported in Table 4. The corresponding estimates of the partial derivatives dS/dz_i and $dS/d\ln x_i$ are reported in Table 5. It will come as no surprise to those of you who analyze data that the least-squares estimates are peculiar in many respects. This collinear data set could hardly support alone the estimation of such a complex model.

A typical discussion of these least-squares estimates might proceed something like the following: "The most significant variable is last year's score, with a surprisingly low coefficient of .21. Students who finish the first grade with a measured reading advantage find that only 21% of the advantage translates into improved scores the following year. Another significant variable is adult education which reduces the effectiveness of teachers, but raises the effectiveness of expenditures so that overall the adult education variable has a positive effect on reading scores.

The three crime variables all have significant coefficients. Crime surprisingly is estimated to have a positive impact on teacher effectiveness. Crime has a negative effect on expenditure effectiveness and overall reduces the rate of learning, though insignificantly so. No other variable in the least-squares estimate in Table 4 has a t-value in excess of 2. Of the derivatives reported in Table 5 only the pupil stability variable has a t in excess of 2, although family income has a t of 1.9. One surprising result in Table 5 is that teachers are estimated to reduce reading scores. Some of these unusual results may be due to the collinearity problem which is solved by stepwise regression.

Three methods were tried, each using SAS default options for termination. The forward selection method includes sequentially the most significant variable until no variable can be found with "significance level" .5 or less. Backward elimination deletes variables sequentially until all included variables have significance level .1 or less. "Stepwise" both adds and removes variables with a significance level .5 to enter and stay. All three methods select last year's score for inclusion. All three methods select the following six variables for exclusion: % Spanish, % Asian, pupil stability, the teacher-pupil ratio interacted with % Spanish and % Black. The derivatives reported in Table 5 are generally not too sensitive to the choice of variables. The exceptions are the teacher/pupil ratio estimated with the backward elimination method. The method that seems to give the best results is the "stepwise" method. It includes few variables all with the right sign. These estimates imply that family income, pupil stability and non-instructional expenditures improve learning rates and blacks reduce learning rates."

I hope you have detected the quotation marks at the beginning and the end of this sequence of sentences. I do not want to be blamed for saying such things. It is clear there is a problem with the least-squares estimates, but you will be lucky indeed if stepwise methods could cure it. Researcher-selection methods are much better because they make use of *a priori* judgments. Actually, in the absence of *a priori* opinion, there is no "collinearity problem" and there is nothing wrong with the least-squares estimates. Remember it was the "wrong" signs that made you want to omit variables. The selection of the "stepwise" estimates over least-squares, forward selection and backward selection was based on judgment about the coefficients. The alternative is a Bayesian analysis which makes use of fully articulated priors in a logical, formally correct manner.

You will find Bayes estimates of the model in Tables 4 and 5. These estimates are based on the prior described above with zero means for every coefficient except the teacher/pupil and the expenditure/pupil coefficients. This prior therefore does not embody much of the information about signs of coefficients alluded to above. As will be shown below, this doesn't matter very much. In any case, the Bayes estimates in Tables 4 and 5 are all sensibly signed with reasonable magnitudes, with the exception of the crimes derivative in Table 5 and the interaction between teachers and crimes and stability in Table 4. The crimes derivative does have a small t-value and it will be further discussed. The Bayes derivatives reported in Table 5 are similar to the least-squares derivatives although the teacher effect is rather close to the prior mean. This is consistent with the low t-values for the teacher variable. The Bayes estimate of the coefficient on last year's score is in contrast only slightly larger than least-squares

and is rather far from the prior mean of one. Only one other of the Bayes estimates is more than two prior standard deviations from the prior mean: the family income derivative. None of the Bayes estimates of the derivatives is more than two sample standard deviations from the least-squares estimates. What is true for this problem, but is not true generally, is that the formal Bayesian analysis has not generated conclusions that are much different than can be obtained by an inspection of the least-squares estimates and standard errors. The construction of the prior was nonetheless useful because it prepared me for the two surprises in this data set: the small coefficient on last year's score and the large derivative for family income.

One issue that could be studied with this data set is the optimal allocation of resources across schools. It could well be that some school environments are so antagonistic to learning that no reasonable efforts could raise reading scores significantly. Estimates for several schools of the change in the reading score induced by a one hundred percent change in the teacher/pupil ratio and the expenditure/pupil ratio are reported in Table 7. School characteristics are reported in Table 6. School #60 is white, wealthy, educated, crime-free and stable. School #66 is stable and Asian. School #215 is Spanish, poor, uneducated and crime-ridden. School #232 is Spanish, poor and uneducated. School #269 is poor and integrated. School #283 is Black, poor, uneducated and crime-ridden. (Notice its surprising improvement of reading scores.) School #308 is integrated and crime-ridden. Least-squares estimates of the effectiveness of teachers and expenditures reported in Table 7 vary unbelievably from school to school. Teachers are estimated to have a very deleterious affect in the white wealthy school, but a strong

positive effect in the poor, black school. This is an example of what happens if you "overfit," that is, if you select a parameterization that is too rich to be supported by the given data. The overfitting problem is cured by the use of a proper prior distribution, and the Bayes estimates vary little from school to school, especially so when compared to their standard errors. The standard errors on the teacher effects are large, and the Bayesian method has done little more than select a sensible point from a large confidence set. The standard errors for the expenditure effects are smaller, and the Bayes estimates are sometimes farther than two sample standard errors from the least-squares estimates, even though the prior for this coefficient is rather diffuse. This occurs often in collinear data sets because prior information about β_1 can have a large impact on the estimate of β_2. It is for this reason that informal inspection of the least-squares results can be very misleading. I have included in Table 7 the stepwise estimates, which also purport to have cured the overfitting problem. In the sense of not varying much from school to school, they are successful. But, to my way of thinking, they overestimate the expenditure effect and underestimate the teacher effect (i.e. zero). As explained in the introduction, this is an expected feature of stepwise procedures.

Everything that has been reported to this point is arbitrary and whimsical. The Bayes estimates and the least-squares estimates are based on two specific prior distributions, neither of which represents my opinions. I certainly don't have the diffuse, uninformed prior which is implicit in least-squares. The Bayes estimates are built on a prior distribution that I selected to approximate my beliefs, but there are an infinity of other distributions that could do as well or better. Accordingly, I will now report

a sensitivity analysis which is intended to show whether inferences hold up to reasonable changes in the prior mean vector and the prior covariance matrix.

Table 8 contains the derivatives of the Bayes estimates with respect to the twenty-four prior means described in Table 3. These derivatives are multiplied by the prior standard errors to show what happens when the prior mean changes by a reasonable amount. Blanks indicate small numbers. Take, for example, the crime variable which has a Bayes estimate with the wrong sign, .22. The derivative with respect to the prior mean on the total Black effect is -.04 in units of the prior standard error. If you are willing to raise the prior Black effect by .5 (one standard error), you can lower the crime effect by .04. I am unwilling to do that because I don't think Blacks increase learning rates. I might raise the prior mean for income or education but that would only make more positive the crime coefficient. I am also willing to have a negative prior mean for crime and a positive mean for stability, both of which reduce the crime estimate, but I am unwilling to change these means by enough to change the sign of the crime estimate. I conclude that I can't change this sign by a reasonable change in the prior means. Generally speaking, the Bayes estimates seem relatively insensitive to choice of the prior mean.

The sensitivity of the coefficient on previous score with respect to the prior variance matrix is reported in Table 9. The upper bound and lower bound for the prior variance matrix are $\sigma_U^2 V$ and $\sigma_L^2 V$ where V is the prior variance defined in Table 3. Theorem 3 is then used to bound the set of estimates from which the extremes are selected. If $\sigma_U^2 = \sigma_L^2$ there is only one covariance matrix $\hat{V}$ that satisfies the bound $\sigma_L^2 V \leq \hat{V} \leq \sigma_U^2 \hat{V}$, namely

$\hat{V} = \sigma_U^2 V$. Then there is only one Bayes estimate. The diagonal of the matrix reported in Table 9 therefore contains various Bayes estimates as the prior covariance is scaled up and down. The lower right is the least-squares estimate, .21. The upper left is the prior estimate, 1, and the center is the Bayes estimate .25. As we move from the center to the upper right, the upper bound for the prior covariance matrix is increasing and the lower bound is decreasing. At the extreme upper right corner, the prior covariance matrix can be anything. In that event, there is a covariance matrix which generates a Bayes estimate as large as 1.04 or as small as .17. It is impossible to get estimates outside that range. These estimates are actually constrained least-squares estimates using linear combinations of the prior constraints.

From my perspective, the family of all prior covariance matrices is much too wide a set of assumptions and I feel confident that I can credibly eliminate many. I suppose I could live with a prior that is twice as diffuse as mine, or twice as sharp, but not too much farther away. In that event, the bound in the box in Table 9 applies. Estimates could range from .23 to .32, a fairly tight interval. What I can learn from the data set is that there is not nearly as much year-to-year retention of reading scores as I had expected. That, unfortunately, is about all that I can learn from this data set.

Bounds for the linear combinations of coefficients are reported in Table 10. Except for the previous score coefficient and the black derivative, these bounds are uselessly wide. For example, if $\sigma_L = \frac{1}{2}$ and $\sigma_U = 2$, the estimate of the teacher coefficient could be as large as 235 or as small as -188. The extreme bound, -1250 to 1260 is absolutely enormous. Similar conclusions apply to all but one of these linear combinations. The black derivative

is barely restricted to be negative, although the lower bound of -.592 is higher than the lower bound for Spanish or Asian. An estimate of -.5 implies that an all black school receives a reading score 50 units less than an all white school. A 50 unit change in y implies a percentile change from 33 to 50 to 66 (see Table 1), which seems like a large difference. The upper bound of -.08 implies an effect which though small is not negligible.

Table 10 should be contrasted with Table 5. As I read Table 5, I conclude from the insensitivity of the estimates to method of estimation that blacks have a perceptibly negative effect on learning and that family income, adult education, pupil stability and expenditures have a perceptibly positive effect. However, as revealed in Table 10, only the black derivative stands up to a sensitivity analysis with respect to the prior covariance matrix. Table 5 is a very inadequate way of reporting the sensitivity of inferences to choice of assumptions, because you see estimates based on only four specific sets of assumptions, three of which are obviously incredible. The column headed $\sigma_L = \frac{1}{2}$, $\sigma_U = 2$ in Table 10 reports estimates based on an infinity of alternative assumptions, most of which are credible to me. An infinity of alternative assumptions are excluded as well, but most of these are incredible to me.

Things may not be quite as bleak as they appear because there are wide ranges of priors that do yield determinate signs. For example, Table 11 reports the bounds for the expenditure effect. Both the lower right and the upper left of this matrix contain only positive elements. Thus if you are willing to state that your prior is certainly much more diffuse than mine

in the sense that your $\sigma_L = 2$, or if you are willing to state that your prior is much more concentrated than mine in the sense that your $\sigma_U = 1/32$, then you can conclude that your estimate of the expenditure effect will be positive. The sign is also positive if $\sigma_L = 1$ and $\sigma_U = 2$. Some of this kind of information is summarized in Table 13. You will notice that both the expenditures and the pupil stability have positive effects for wide ranges of priors, in particular those with $\sigma_L = 2$.

Estimates and bounds for the third and fourth grade data are reported in Tables 12 and 13. The general conclusions are the same as before: the retention of reading percentiles from year-to-year is small; schools with large percentages of Blacks have lower rates of accumulation of reading skills. (This latter conclusion does not apply to the third grade.) Incidentally, I would have expected the coefficient on last years score to increase with the grade level, but the opposite seems to be the case.

4.0 Conclusion

My primary purpose has been to convince you that Bayesian methods can profitably be used to study data sets. The principal road blocks that have prevented Bayesian methods from being used in the past have been the cost to the researcher of forming a prior distribution and the incredulity that greets a prior once it has been formalized. Both of these road blocks are substantially eliminated by an analysis indicating the sensitivity of inferences to choice of prior. When the inferences are reasonably insensitive to choice of prior, you need spend little effort precisely defining the prior and you need be little concerned that any particular distribution is incredible. When the inferences are sensitive to choice of prior, as I have demonstrated for the reading score data, you are forced to find ways to define more sharply your prior distribution, or you must obtain better data, or you must suspend inference. In the case of the reading scores I have no other data available, and since I am not an expert on the subject I have only vaguely held prior judgements and am unable to define more precisely my prior. I, unhappily but <u>honestly</u>, admit that there is little I have learned from this data set. I have learned that the retention of reading success from year-to-year is much lower than I expected, and I conclude that there is a large schooling value-added. I have also learned that schools with large proportions of Blacks have somewhat lower rates of learning. But that is all.

The alternative to the Bayesian method, which I have suggested might have led to the stepwise estimates in Table 4, apparently yields sharp results. But a data analyst who selected these estimates for reporting purposes is either fooling himself, fooling his clients or both.

There are two important shortcomings of this analysis of reading scores. First it is estimation oriented. I have produced bounds for the estimated Black effect, but have not attached measures of uncertainty to these bounds. What would be more interesting is bounds for the probability that the Black effect is less than some number. When the estimates are limited to be less than zero, you can conclude that the probability that the coefficient is less than zero is greater than one-half. This is not particularly useful.

The other shortcoming is the dependence on the assumption of normality for both the prior and the sampling process. There do not yet exist computer routines for dealing with non-normal priors but there are many "data-analytic" methods for dealing with non-normal sampling processes. These methods can produce "robust" estimates of the coefficients and covariance matrix which can then be used as input into a Bayesian analysis. Although this is not a fully correct procedure, I doubt that the distortion is great. I am worried in particular about the improvement in the reading scores for school #308 and I wonder how this and other unusual schools affect the least-squares estimates, and the subsequent Bayesian analysis.

TABLE 1

Reading scores S and Percentiles P

$S = 100P/(100-P)$

Equal changes in ln S

S	P
12.5	11
25	20
50	33
100	50
200	66
400	80
800	89
1600	94
3200	96.9

Equal changes in S

S	P
50	33
100	50
150	60
200	66
250	71
300	75
350	78
400	80

TABLE 2

Variables

P = COOP reading percentile score of grade i, Spring 1973.

P_0 = COOP reading percentile score of grade i-1, Spring 1972.

x_1 = Teacher-pupil ratio, Fall 1972.

x_2 = Non-instructional expenditures ($'s) per pupil, special plus regular district funds, 1972-1973.

z_0 = Constant.

z_1 = % Spanish surname pupils, Fall 1972.

z_2 = % Black pupils, Fall 1972.

z_3 = % Asian pupils, Fall 1972.

z_4 = Logarithm of family income, 1970 Census.

z_5 = Adult educational attainment: $(100)^{-1}$[6(% attaining 6th grade) + 9(% attaining 7th to 11th grade) + 13(% attaining grade 12 to Jr. Coll.) + 16(% attaining college diploma)].

z_6 = Total crimes per 1,000 pupils (robberies, assaults, sex offenses, burglaries, thefts, vandalism, arson, narcotics, loitering and trespassing), 1972-1973.

z_7 = Pupil stability (percent of students who stay in school from fall to spring), 1972-73, $0 < z_7 < 100$.

Market value of homes.

% funds spent on instruction.

Student body income.

Parent volunteers.

% pupils on free lunch.

% pupils on AFDC.

% Spanish surname teachers.

% Black teachers.

% Asian teachers.

% American Indian teachers.

Pupil school-year transiency (leavers + arrivers/enrollment)

% American Indian pupils.

Books in library.

Professional books in library.

Eight MM films.

Film strips.

Recordings.

Teacher transfer requests.

Certified teachers.

% Absent.

TABLE 3

Prior Information

Parameter	Mean	2(Stnd. Error)	Explanation
δ	1	.2	Scores vary little over short intervals.
β_0	0	∞	The constant depends on functional form approximation about which little is known.
dS/dz_1	0	1	When the percent Spanish varies by 1, scores vary by $\pm$ 1.
dS/dz_2	0	1	When the percent Black varies by 1, scores vary by $\pm$ 1.
dS/dz_3	0	1	When the percent Asian varies by 1, scores vary by $\pm$ 1.
dS/dz_4	0	50	When family income changes by 10%, scores vary by $\pm$ 5.
dS/dz_5	0	10	When adult education varies by one grade, scores vary by $\pm$ 10.
dS/dz_6	0	5	When crimes per 1,000 pupils varies by 1, scores vary by $\pm$ 5.
dS/dz_7	0	1	When the percent stable increases by 1, scores vary by $\pm$ 1.
$dS/d\ln x_1$	25	50	When teacher/pupil increases by 10%, scores increase by 2.5 $\pm$ 5.
$dS/d\ln x_2$	12.5	50	When expend./pupil increases by 10%, scores increase by 1.25 $\pm$ 5.
θ_1	0	.1	.1 times standard error for dS/dz_i (See text)
θ_2	0	.1	
θ_3	0	.1	
θ_4	0	5	
θ_5	0	1	
θ_6	0	.5	
θ_7	0	.1	

TABLE 3 (Continued)

Parameter	Mean	2(Stnd. Error)	Explanation
γ_1	0	.05	
γ_2	0	.05	
γ_3	0	.05	
γ_4	0	2.5	.05 times standard error for dS/dz_i (See text)
γ_5	0	.5	
γ_6	0	.25	
γ_7	0	.5	

TABLE 4

Individual Coefficient Estimates (Absolute t-values) - Second Grade

Variable	Least Squares		Forward Selection (.5 sig. to enter)		Backward Elimination (.1 sig. to stay)		"Stepwise" (.5 to enter and stay)		Bayes
Last Year's Score	.21	(13)	.22	(13)	.22	(14)	.23	(14)	.25
Constant	3417	(.9)	7054		-253		-634		-417
% Spanish	-5.7	(.8)							-.38
% Black	-5.2	(1)			-.22	(1.9)			-.31
% Asian	1.5	(.1)							-.24
Family Income	-48	(.1)	-759	(3)			70	(4.4)	50
Adult Education	-198	(1.2)							4.4
Crime/1,000 Pupils	36	(3.8)			40.5	(6)			.99
Pupil Stability	-10	(.9)							.31
Teacher/Pupil Times									
Constant	930	(.9)	2038	(3)					24.3
% Spanish	-1.3	(.7)							.002
% Black	-1.2	(.7)							.004

Table 4 (continued)

Variable	Least Squares		Forward Selection		Backward Elimination		"Stepwise"		Bayes
% Asian	1	(.3)	.56	(.9)					.001
Family Income	-17	(.1)	-221	(3)					-.25
Adult Education	-49	(1)			-3.3	(2.8)			-.03
Crime/1,000 Pupils	8.4	(2.9)			9.7	(4.6)			.09
Pupil Stability	-3.2	(1)			-.4	(3)			-.001
Expenditures/Pupil									
Times Constant	-203	(1)	-191	(1.7)	-115	(3.2)			-3.8
% Spanish	.26	(.8)	-.05	(1.1)			-.09	(2.7)	-.004
% Black	.23	(.9)	-.07	(2)			-.05	(1.8)	-.002
% Asian	.3	(.5)	.32	(.8)					.0006
Family Income	9.1	(.3)	16.6	(1.3)	14.7	(4)			.26
Adult Education	10.1	(1)	2.1	(1.5)					.04
Crimes/1,000 Pupils	-1.8	(3)			-1.9	(3.6)			-.1
Pupil Stability	.17	(.2)	.26	(2.6)			.2	(3.5)	.11
R^2	.654		.635		.649		.621		
R.M.S.E.	62.4		63.1		61.8		63.8		

TABLE 5

Estimated Derivatives Evaluated at Data Means - Second Grade
(Absolute t-values in parenthesis)

Variable	Least Squares		Forward Selection	Backward Elimination	"Stepwise"	Prior	Bayes	
% Spanish	-.206	(.75)	-.235		-.444	0	-.4	(2.4)
% Black	-.29	(1.5)	-.33	-.215	-.256	0	-.34	(2.4)
% Asian	-.309	(.7)	-.275			0	-.23	(.7)
Family Income	50.8	(1.9)	42.4	71	69.9	0	51.8	(4.1)
Adult Education	10	(1.3)	10.2	10.7		0	4.7	(1.2)
Crimes/1,000	-.235	(.4)		-.255		0	.22	(.5)
Pupil Stability	1.27	(2.5)	1.27	1.3	.957	0	.86	(2.6)
Teachers/Pupils	- 12	(.3)	-22.3	12.3		25	22.3	(1.1)
Expenditures/Pupils	7.5	(1.4)	7.32	-131.5	12.2	12.5	6.9	(1.4)

TABLE 6

Characteristics of Selected Schools

School	#60	#66	#215	#232	#269	#283	#308	Avg.
% Spanish	2.5	20.9	80.8	99.1	38.4	.3	25.3	25.9
% Black	0	.5	14.9	0.0	28.2	99.7	40.3	21.3
% Asian	1.3	75.8	.2	0.0		0.0	23.7	3.6
Log Family Income	10.6	9.3	8.9	9.1	8.8	8.6	9.4	9.3
Family Income	39,604	11,173	7,251	8,714	6,332	5,469	12,449	12,212
Adult Education	14.10	10.1	8.7	8.3	9.9	9.9	11.6	11.3
Crimes/1,000 Pupils	3.2	8.8	24.3	13.0	8.2	58.0	26.0	8.3
Pupil Stability	88.2	92.0	79.8	83.9	76.7	78.7	76.3	79.1
Pupil/Teacher	31.4	23.4	22.2	20.5	22.2	18.4	28.0	26.6
Expenditures/Pupil	500	102	112	221	227	87	283	158
Reading Percentile								
First Grade 1973	66	16	38	54	30	68	37	52.2
First Grade 1972	89	20	24	65	35	37	54	50.6
Second Grade 1973	77	32	28	51	29	78	30	46.9
Second Grade 1972	62	48	25	37	31	32	19	45.9
Third Grade 1973	78	32	27	26	17	48	26	41.5
Third Grade 1972	75	28	25	13	14	19	30	39.8
Fourth Grade 1973	79	31	29	23	16	25	34	41

TABLE 7

Estimates of Teacher and Expenditure Effects for Several Schools - Second Grade
(Standard errors in parenthesis)

School	Estimates of Teacher Effect				Estimates of Expenditure Effect			
	Least Squares		Bayes		Least Squares		Bayes	
#60	-189	[128]	21.3	[21.1]	46.9	[20]	9	[6]
#66	112	[218]	22.3	[21]	12.2	[43]	8.4	[6]
#215	186	[72]	24	[21]	-40.5	[16]	4.9	[5]
#232	88.1	[87]	22.9	[21]	-20.2	[19]	6.5	[5]
#269	43.6	[58]	22.6	[21]	-8.05	[10]	6.4	[5]
#283	428	[134]	27.2	[24]	-95.4	[29]	1.5	[8]
#308	124	[99]	23.9	[21]	10.7	[18]	4.9	[5]
AVG.	-12	[48]	22.3	[21]	7.5	[5]	6.9	[5]

Estimates of Expenditure Effect
"Stepwise" - Teacher Effect Is Zero

17.3
16.3
7.6
7.5
10.2
10.3
10.7
12.2

TABLE 8

Changes in Posterior Means Induced by One Standard Deviation Change in Prior Mean - Second Grade
[Bayes Estimates in Parenthesis]

Prior On	Stnd. Err.	Spanish[1]	Black[1]	Asian[1]	Income[2]	Education[3]	Crimes[1]	Stability[1]	Teachers[3]	Expenditures[3]
		[-.4]	[-.34]	[-.23]	[51.8]	[4.7]	[.22]	[.86]	[22.3]	[6.9]
Previous Score	.1				-1			-.01	-.1	-.1
Spanish	.5	.06	.03		1	.6			-1.4	
Black	.5	.03	.04		1	.1	-.04		-1.5	.2
Asian	.5			.2	1		.02	-.01	.8	-.3
Income	25	.02	.02	.01	8	-.1	.02	-.04	1.4	-.2
Education	5	.06	.01		-5	2.9	.04	.01	1.8	-.2
Crimes	2.5		-.01			.1	.08		-.5	-.2
Stability	.5			-.01	-2	.1	-.01	.22	-.6	.2
Teachers	25	-.03	-.03	.02	1	.4	-.05	-.01	17.1	-.7
Teachers Times:										
Spanish	.05								-.2	
Black	.05	.01					-.01		-.3	
Asian	.05									

TABLE 8 (Continued)

Prior On	Stnd. Err.	Spanish[1]	Black[1]	Asian[1]	Income[2]	Education[3]	Crimes[1]	Stability[1]	Teachers[3]	Expenditures[3]
Income	2.5								-.2	
Education	.5								.2	
Crime	.25						-.03		-.1	
Stability	.05									
Expenditures	25			-.01			-.02		-.7	.9
Expenditures Times:										
Spanish	.025						.01		-.2	-.1
Black	.025		.01				-.02		.2	.1
Asian	.025									
Income	1.25									
Education	.25									
Crime	.125						-.04		.2	
Stability	.25									

[1]Blank less than .01; [2]Blank less than 1; [3]Blank less than .1.

TABLE 9

Bounds for Coefficient on Previous Score - Second Grade

Prior Variance Matrix Bounded between $\sigma_L^2 V$ and $\sigma_U^2 V$,

Where V Is the Selected Variance Matrix

σ_L \ σ_U	0	1/32	1/16	1/8	1/4	1/2	1	2	4	8	16	32	∞
0	1.00 1.00	1.00 .96	1.00 .88	1.01 .68	1.01 .45	1.02 .30	1.02 .23	1.02 .21	1.03 .20	1.03 .19	1.03 .19	1.03 .18	1.04 .17
1/32		.97 .97	.97 .88	.97 .68	.98 .45	.98 .30	.99 .23	.99 .21	.99 .20	.99 .19	.99 .19	1.00 .18	1.00 .18
1/16			.88 .88	.88 .69	.89 .45	.90 .30	.90 .23	.90 .21	.90 .20	.90 .20	.91 .19	.91 .19	.92 .18
1/8				.69 .69	.69 .45	.70 .30	.70 .24	.71 .22	.71 .21	.71 .20	.71 .20	.71 .19	.72 .18
1/4					.46 .46	.46 .31	.47 .25	.47 .22	.47 .21	.47 .21	.47 .20	.48 .20	.48 .19
1/2						.31 .31	.32 .25	.32 .23	.32 .22	.32 .21	.32 .21	.32 .21	.33 .20
1							.25 .25	.26 .23	.26 .22	.26 .22	.26 .21	.26 .21	.26 .20
2								.23 .23	.23 .22	.24 .22	.24 .21	.24 .21	.24 .21
4									.23 .23	.23 .22	.23 .22	.23 .21	.23 .21
8										.22 .22	.22 .22	.22 .21	.23 .21
16											.22 .22	.22 .21	.22 .21
32												.22 .22	.22 .21
∞													.21 .21

TABLE 10

Estimates and Bounds for Derivatives - Second Grade

	Prior	Bayes	Least-squares	Bounds $\sigma_L = \frac{1}{2}, \sigma_U = 2$		Bounds $\sigma_L = 0, \sigma_U =$	
				Lower	Upper	Lower	Upper
Previous Score	1	.25	.21	.228	.318	.174	1.04
% Spanish	0	-.4	-.21	-1.53	.764	-7.27	7.06
% Black	0	-.33	-.29	-.592	-.08	-5.15	4.86
% Asian	0	-.23	-.31	-2.61	2.17	-12.4	12.0
Family Income	0	51.8	50.8	-75.8	167	-686	736
Adult Education	0	4.7	10	-33.9	42.9	-200	210
Crimes/1,000	0	.22	-.23	- 1.6	1.91	-16.5	16.2
Pupil Stability	0	.85	1.27	-1.89	3.49	-12.6	13.8
Teachers/Pupils	25	22.3	- 12	-188	235	-1250	1260
Expenditures/Pupils	12.5	6.9	7.5	-10.2	24.1	-131	151

Bounds for the Estimate of the Expenditure Derivative Prior Variance Matrix Bounded between $\sigma_L^2 V$ and $\sigma_U^2 V$ - Second Grade

σ_L \ σ_U	0	1/32	1/16	1/8	1/4	1/2	1	2	4	8	16	32	∞
0	12.5	16.6	27.3	54.9	94.6	122.	135.	141.	145.	147.	149.	150.	151.
	12.5	7.49	-5.09	-35.6	-75.7	-103.	-115.	-122.	-126.	-127.	-129.	-130.	-131.
1/32		11.6	22.3	49.8	89.4	117.	129.	136.	140.	142.	143.	144.	146.
		11.6	-.962	-31.4	-71.4	-98.4	-111.	-118.	-121.	-123.	-124.	-125.	-126
1/16			9.68	37.2	76.5	104.	116.	123.	126.	128.	130.	131.	132.
			9.68	-20.7	-60.4	-87.2	-99.6	-106.	-109.	-111.	-112.	-113.	-115.
1/8				6.80	45.8	72.8	84.8	90.9	94.4	96.4	97.6	98.4	99.7
				6.80	-32.6	-59.0	-71.1	-77.2	-80.5	-82.0	-83.1	-84.0	-85.4
1/4					6.43	33.1	44.9	50.7	53.9	55.7	56.9	57.6	58.7
					6.43	-19.8	-31.6	-37.4	-40.4	-41.8	-42.8	-43.6	-44.8
1/2						6.96	18.6	24.1	27.0	28.7	29.8	30.5	31.4
						6.96	-4.76	-10.2	-13.0	-14.2	-15.1	-15.9	-17.0
1							6.90	12.2	14.9	16.5	17.5	18.1	19.0
							6.90	1.64	-.850	-2.07	-2.93	-3.63	-4.56
2								6.91	9.56	11.2	12.2	12.8	13.6
								6.91	4.48	3.26	2.40	1.71	.817
4									7.12	8.74	9.77	10.4	11.2
									7.12	5.91	5.05	4.35	3.46
8										7.52	8.55	9.15	9.93
										7.52	6.66	5.97	5.09
16											7.69	8.28	9.03
											7.69	7.01	6.16
32												7.60	8.34
												7.60	6.76
∞													7.50
													7.50

TABLE 12

Estimates and Bounds, Third and Fourth Grades
Bounds Based on $\sigma_L = \frac{1}{2}$, $\sigma_U = 2$

		Third Grade				Fourth Grade			
	Prior	Bayes	Least-squares	Bounds		Bayes	Least-squares	Bounds	
Previous Score	1	.39	.23	.292	.595	.183	.116	.152	.277
% Spanish	0	-.31	.41	-1.34	.90	-.63	-.5	-1.23	.02
% Black	0	-.51	-.69	-1.26	.36	-.52	-.59	-.913	-.092
% Asian	0	-.09	-.22	-2.43	2.23	.13	.3	-1.09	1.35
Family Income	0	27.3	-17.8	-91.9	.29	75.2	63.4	-2.08	142
Adult Education	0	8.43	36.2	-24.1	43.2	3.8	11.6	-21.2	29.1
Crimes/1,000	0	.75	1.3	-1.18	2.55	-.04	-.33	-1.1	.944
Pupil Stability	0	.37	1.03	-2.13	3.02	.63	.25	-.92	2.03
Teachers/Pupils	25	12.5	-85.6	-171	198	73.6	28.8	-68.5	208
Expenditures/Pupils	12.5	5.7	8.4	-12.9	24.4	69.6	-293	-5.68	11.6

TABLE 13

Conditions under which the Sign Is Determined

	Second Grade		Third Grade		Fourth Grade	
	sign	condition (σ_U and σ_L)	sign	condition (σ_U and σ_L)	sign	condition (σ_U and σ_L)
Previous Score	+	≥ 0	+	≥ 0	+	≥ 0
% Spanish	-	≥ 16	+	≥ 16	-	≥ 2
% Black	-	$\geq .4$	-	≥ 2	-	≥ 1
% Asian	-	≥ 8			+	≥ 16
Family Income	+	≥ 4			+	≥ 2
Adult Education	+	≥ 8	+	≥ 4	+	≥ 4
Crimes/1,000	-	≥ 32	+	≥ 8	-	≥ 16
Pupil Stability	+	≥ 2	+	≥ 4	+	≥ 4
Teachers/Pupils	+	$\leq 1/16$	+/-	$\leq 1/8/\geq 32$	+	$\leq 1/16$ or ≥ 8
Expenditures/ Pupils	+	$\leq 1/32$ or ≥ 2	+	$\leq 1/8$ or ≥ 4	+/-	$\leq 1/16/\geq 16$