REMARKS ON THE ESTIMATION OF COEFFICIENTS OF A REGRESSION

IN THE PRESENCE OF UNKNOWN EXPLANATORY VARIABLES

Herman Chernoff*

Massachusetts Institute of Technology

and

Harvard University

In the linear regression model $Y = \beta_1 X_1 + \beta_2 X_2 + u$, the coefficients β_1 and β_2 may be estimated by least squares. If the explanatory variable X_2 is not observed, the regression of Y on X_1 will give an estimate of β_1 whose bias will depend on the correlation between X_1 and X_2. However qualitative knowledge about X_2 can be exploited. We treat the case where the known and unknown explanatory variables and the coefficients are nonnegative and where it is known that for some, but not which, data points, the unknown explanatory variables are relatively small.

1. Introduction.

A source of difficulty in estimating the effect of one variable on another, especially in observational studies, is that the explanatory model may omit a causal variable. Under some circumstances, this difficulty may be serious. If the omitted variable is unimportant, i.e. it has a relatively small effect, it may be safe to ignore it. If it is uncorrelated with the other explanatory variables, it may also be ignored in linear regression models. If it is correlated with the explanatory variables, and one desires only to use these for prediction, one may proceed without it, as long as that correlation is

* Research partially supported by NSF grant #MCS82-01732.

AMS 1980 subject classification. 62J05, 62H30.

Key words and phrases. Linear regression, missing variables, causal models, correlation, cluster, mode.

to be kept constant, the effect of the known independent variables can be assessed.

In other cases, standard regression analysis, ignoring the omitted variable, can lead to important errors, including wrong signs and a host of problems associated with the term spurious correlation.

In a recent paper, Rutan and Brown (1984) propose a method of dealing with such an estimation problem, in the context of analytical chemistry applications, by using adaptive Kalman filters to compensate for low quality models. While the use of Kalman filters may be relevant to these particular applications, one is led to raise the fundamental question of what is the basic principle that can be used to avoid the classical dilemma.

To omit a causal variable or to say that it's value is not known is not the same as to say that nothing is known about it. In the above applications there are several known facts. The unknown causal variables and their effects are known to be nonnegative. Moreover, it is known, or assumed, that there is a nontrivial, but unspecified set of data points for which the unknown variables and their effects are negligible.

The Kalman filter approach exploits still more information. It uses the fact that there is a natural (time) ordering of the data, that the unknown variables are relatively unimportant at the early times, and their effect is a smooth function of time.

In this discussion, I propose to omit these latter assumptions and confine attention to the positivity and occasional negligibility. The moral is that quantitatively vague background information can sometimes be exploited under circumstances where the lack of such information leaves one helpless. Needless to say, especially at this occasion, this moral has been effectively demonstrated by others.

2. Two regression models and modal estimates.

We consider two regression models. These are

$$Y_i = \beta_1 X_{i1} + V_i + u_i, \qquad i=1,2,\ldots,n \tag{2.1}$$

and

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + V_i + u_i, \qquad i=1,2,\ldots,n, \tag{2.2}$$

where Y_i, X_{i1} and X_{i2} are observed, β_1 and β_2 are unknown, and the residuals u_i are assumed to be normally distributed, with mean 0 and constant variance σ^2, independent of each other and of the other variables. The variables V_i are not observed. They represent the effect of the unknown causal variable and may be correlated with X_{i1} and X_{i2}.

The standard regression of Y on X_1 for (2.1), ignoring V, would yield an estimate of β_1 which is approximately

$$\sigma_{XY_1}/\sigma^2_{X_1} = \beta_1 + \sigma_{VX_1}/\sigma^2_{X_1} .$$

If V is correlated with X_1, the estimate could be seriously biased. That is also the case for the model (2.2).

To help fix our notions let us consider an example which is similar to that appearing in analytical chemistry. Let

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i, \quad i=1,2,\ldots,n \tag{2.3}$$

where

$$X_{ij} = \exp\{-(\tfrac{i}{n} - \mu_j)^2/2\sigma_j^2\} \quad j=1,2,3;\ i=1,2,\ldots,n \tag{2.4}$$

and μ_j and σ_j^2 are specified constants. For example, we have experimented with the case $n = 200$, $\beta_1 = \beta_2 = \beta_3 = 1$, $\mu_1 = 0.2$, $\mu_2 = 0.5$, $\mu_3 = 0.75$, $\sigma^2 = \sigma_1^2 = \sigma_3^2 = 0.01$, and $\sigma_2^2 = 0.0125$. For model (2.1) we can take $V_i = \beta_2 X_{i2} + \beta_3 X_{i3}$ or $\beta_1 X_{i1} + \beta_2 X_{i2}$ or $\beta_1 X_{i1} + \beta_3 X_{i3}$. For model (2.2) we can take $V_i = \beta_1 X_{i1}$ or $\beta_2 X_{i2}$ or $\beta_3 X_{i3}$. We shall not use the fact that V_i is a reasonably smooth function of i. However, we do wish to use the facts that there is a substantial set of i for which $\beta_1 X_{i1}$ is much larger than V_i which, in turn, is smaller than σ, and β_1, X_{i1}, and V_i are all nonnegative.

For model (2.1), we use the i-th observation to estimate β_1 with

$$\hat{\beta}(i) = X_{i1}^{-1} Y_i = \beta_1 + X_{i1}^{-1}(V_i + u_i) > \beta_1 + X_{i1}^{-1} u_i, \tag{2.5}$$

and for some i, $X_{i1}^{-1} V_i$ is small compared to β_1, and the last inequality is almost an equality. If the $X_{i1}^{-1} Y_i$ are used as estimates of β_1 they will generally be biased positively, but some of them will cluster close to β_1. Thus, one may expect the sample distribution of $\hat{\beta}(i)$ to have a mode close to, but above, β_1. Equivalently, the $\hat{\beta}(i)$ should have a cluster centered close to β_1. The magnitude of the X_{i1} when V_i is small will affect the spread of the points. Thus, it seems natural to weight the $\hat{\beta}(i)$. In particular, let $w_i = X_{i1}^2$ and let

$$s(\beta) = \sum_{\hat{\beta}(j) \leq \beta} w_j \tag{2.6}$$

represent the cumulated weight for β_1 no larger than β. When $s(\beta)$ is plotted against β, the modal value corresponds to that value of β for which s increases most rapidly. In particular, one way of defining this mode is to select an interval length δ and to let $\hat{\beta}_1$ be that value of β for which $s(\beta + \delta/2) - s(\beta + \delta/2)$ is a maximum. There may be several modes. In that case, one close to the minimum value of $\hat{\beta}(i)$ is recommended. The reader should note that the above is somewhat short of a formal definition.

The above modal estimate $\hat{\beta}$, is likely to be positively biased. We shall not analyze its properties, but plan to use it for a first approximation. To help generalize this estimate for model (2.2) it is convenient to interpret it in the following manner. For each estimate $\hat{\beta}(i)$ construct a kernel function

$$K(\beta, \hat{\beta}(i)) = K^*(\delta^{-1}[\beta - \hat{\beta}(i)]) \tag{2.7}$$

where $K^*(z) = 0$ for $|z| > 0.5$ and 1 for $|z| \leq 0.5$.

Then $\hat{\beta}$ is the value of β which provides the appropriate local maximum of

$$g(\beta) = \Sigma w_i K(\beta, \hat{\beta}(i)). \tag{2.8}$$

One way of representing the data in model (2.2) is to draw the line $Y_i = X_{i1}\beta_1 + X_{i2}\beta_2$ in the (β_1,β_2) plane for each point (Y_i, X_{i1}, X_{i2}). The positivity of $\beta_1, \beta_2, X_{i1}, X_{i2}$, and V_i suggests that the true $\underline{\beta} = (\beta_1,\beta_2)$ will usually lie somewhere below and to the left of that line. Our assumptions about the V_i imply that $\underline{\beta}$ should lie close to some of these lines. Thus, we have a reasonable expectation of finding $\underline{\beta}$ close to a densely populated corner of the convex intersection of the sets below these lines. Some vertical adjustment to allow for the effect of σ is appropriate.

A more direct generalization of the one dimensional modal estimate follows. Two points (Y_i, X_{i1}, X_{i2}) and (Y_j, X_{j1}, X_{j2}) determine two lines, as above, which intersect at

$$\hat{\beta}_1(i,j) = \frac{X_{j2}Y_i - X_{i2}Y_j}{X_{j2}X_{i1} - X_{i2}X_{j1}}, \quad \hat{\beta}_2(i,j) = \frac{-X_{j1}Y_i + X_{i1}Y_j}{X_{j2}X_{i1} - X_{i2}X_{j1}}.$$

$$\hat{\beta}_1(i,j) = \beta_1 + D^{-1}[X_{j2}(V_i + u_i) - X_{i2}(V_j + u_j)] \tag{2.9}$$

and

$$\hat{\beta}_2(i,j) = \beta_2 - D^{-1}[X_{j1}(V_i + u_i) - X_{i1}(V_j + u_j)] \tag{2.10}$$

where

$$D = X_{j2}X_{i1} - X_{i2}X_{j1} = D(i,j). \tag{2.11}$$

Then $\hat{\underline{\beta}}(i,j) = (\hat{\beta}_1(i,j), \hat{\beta}_2(i,j))$ has covariance matrix

$$(2.12) \qquad \Sigma(i,j) = \frac{\sigma^2}{D^2} \left\| \begin{matrix} X_{i2}^2 + X_{j2}^2 & -(X_{i1}X_{i2} + X_{j1}X_{j2}) \\ -(X_{i1}X_{i2} + X_{j1}X_{j2}) & X_{i1}^2 + X_{j1}^2 \end{matrix} \right\|$$

with inverse $\sigma^{-2}J(i,j)$ where

$$(2.13) \qquad J(i,j) = [\underline{X}_i\underline{X}_i' + \underline{X}_j\underline{X}_j']$$

and $\underline{X}_i' = (X_{i1}, X_{i2})$. Note that

$$(2.14) \qquad \det J(i,j) = D^2.$$

Now let $K(\underline{\beta}, \hat{\underline{\beta}}(i,j)) = 1$ for the ellipse, of area $\pi\delta^2$, defined by

$$(2.15) \qquad [\underline{\beta} - \hat{\underline{\beta}}(i,j)]'J(i,j)[\underline{\beta} - \hat{\underline{\beta}}(i,j)] < |D|\delta^2,$$

and let $K = 0$ outside that ellipse. We then seek the local maximum of

$$(2.16) \qquad g(\underline{\beta}) = \sum_{i \neq j} D^2(i,j)K(\underline{\beta}, \hat{\underline{\beta}}(i,j)).$$

to be our modal estimate $\hat{\underline{\beta}}$.

3. <u>The model with one known explanatory variable</u>.

The investigation of data from artificial examples suggests that if σ is small enough and there are enough points for which V is negligible, the plot of Y versus X_1 would ordinarily provide a reasonably sharp estimate of β_1 without much recourse to formal theory or procedures. However, an all purpose theoretical procedure is difficult to develop if we lack large samples and require robustness.

A strategy which might apply if the sample size were large and the model were reliable, is to estimate the cumulant generating function of

$$Y - \beta X_1 = (\beta_1 - \beta)X_1 + V + u$$

and subtract that of u, i.e. $t\sigma^2/2$, and estimate the distribution of $(\beta_1 - \beta)X_1 + V$ (as well as σ^2). For $\beta > \beta_1$, that distribution would assign positive probability to negative values. Thus, β_1 would correspond to the smallest value of β for which $(\beta_1 - \beta)X_1 + V$ has zero probability of being negative. The implementation of this strategy is likely to lead to rules which require large samples and may lack robustness.

We outline an alternative approach. For the sake of simplicity, we shall assume that σ is known. Incidentally, if there were a large set of points for which $\beta_1 X_{i1} + V_i$ were small, we could use those data to estimate σ.

Our approach consists of assuming, in a limited sense, that the distribution of V is the mixture of 0 with probability p and a uniform random variable on $(0,\tau^{-1})$ with probability $(1-p)$. Then, by identifying certain properties of the sample with corresponding properties of the distribution of $Y - \beta X_1$ for trial values of β, estimates of β_1 may be derived which are moderately robust.

Under the above assumption the distribution of

$$Z = V + u \tag{3.1}$$

is given by the density

$$f(z) = \frac{p}{\sigma}\,\phi\left(\frac{z}{\sigma}\right) + (1-p)\tau\left\{\Phi\left(\frac{z}{\sigma}\right) - \Phi\left(\frac{z - \tau^{-1}}{\sigma}\right)\right\} \tag{3.2}$$

where ϕ and Φ are the standard normal density and cumulative distribution functions. Let

$$z^* = z - \tau^{-1}. \tag{3.3}$$

Then, the cumulative distribution function of Z is

$$F(z) = p\Phi(\tfrac{z}{\sigma}) + (1 - p)\tau\sigma[\psi(\tfrac{z}{\sigma}) - \psi(\tfrac{z^*}{\sigma})] \tag{3.4}$$

where

$$\psi(v) = \phi(v) + v\Phi(v). \tag{3.5}$$

We also note that

$$f'(z) = \sigma^{-2}\{-p\tfrac{z}{\sigma}\,\phi(\tfrac{z}{\sigma}) + (1 - p)\tau\sigma[\phi(\tfrac{z}{\sigma}) - \phi(\tfrac{z^*}{\sigma})]\} \tag{3.6}$$

and, defining

$$G(z) = \int_{-\infty}^{z} vf(v)dv,$$

$$G(z) = \sigma\{-p\phi(\tfrac{z}{\sigma}) + (1-p)\tau\sigma[\zeta(\tfrac{z}{\sigma})] - \zeta(\tfrac{z^*}{\sigma})] - (1-p)\psi(\tfrac{z^*}{\sigma})\}, \tag{3.7}$$

where

$$\zeta(v) = \frac{1}{2}\{v\phi(v) + (v^2 - 1)\Phi(v)\}. \tag{3.8}$$

Note that if $\tau\sigma$ is small,

$$f'(z) = 0 \implies \frac{z}{\sigma} \approx \frac{1-p}{p}\tau\sigma \equiv u_0. \tag{3.9}$$

Thus, under this model, the distribution of $Y_i - \beta_1 X_{i1} = Z_i$ has its mode at $z \approx \sigma u_0 > 0$.

To return to the task of estimating β_1, we shall select trial values β of β_1, and see which is, in a limited sense, most consistent with our model. For this case of known σ, we use

$$m_0(\beta) = \sum_{i=1}^{n} \chi(Y_i < \beta X_{i1}) \tag{3.10}$$

$$m_1(\beta) = \sum_{i=1}^{n} \chi(Y_i < \beta X_{i1} + \sigma) - m_0(\beta) \tag{3.11}$$

and

$$m_2(\beta) = \sum_{i=1}^{n} \chi(Y_i < \beta X_1 + 2.6\sigma) - m_0(\beta) - m_1(\beta) \tag{3.12}$$

where $\chi(E)$ is the characteristic function of the event E. Then, assuming $\tau\sigma$ is small, and hence $\phi(z^*/\sigma)$ and $\psi(z^*/\sigma)$ can be ignored for $z/\sigma = 0,1$, and 2.6,

$$\begin{aligned} E[m_0(\beta)] &= nF(0) \approx 0.5000\,np + 0.3989n\sigma\tau(1-p), \\ E[m_1(\beta)] &= n[F(\sigma) - F(0)] \approx 0.3413np + 0.6843n\sigma\tau(1-p), \\ E[m_2(\beta)] &= n[F(2.6\sigma) - F(\sigma)] \approx 0.1540np + 1.5181n\sigma\tau(1-p). \end{aligned} \tag{3.13}$$

Two additional considerations enter in exploiting equations (3.13). First, points for which X_{i1} is small will contribute little to the change in $Y - \beta X_{i1}$ as β varies, and will accomplish little in permitting us to discriminate between good and bad approximations to β_1. Thus, we shall confine attention to points where $X_{i1} > c$ for some suitable constant c. Second, our model for the distribution of V is unrealistic. It may fit well in the neighborhood of $V = 0$. More precisely, the conditional distribution of V, for $V < 3\sigma$, may resemble that of our model. Thus, it is important to distinguish between the original sample size n, or even the sample size truncated by the restriction $X_{i1} > c$, and an effective sample size that fits the data to our model for $V < 3\sigma$.

Given the value of β_1 we can estimate $n^* = np$ and $\rho = n\sigma\tau(1-p)$ from

$$\text{(3.14)} \qquad \begin{aligned} m_1(\beta_1) &= 0.3413\hat{n}^* + 0.6843\hat{\rho} \\ m_2(\beta_1) &= 0.1540\hat{n}^* + 1.5181\hat{\rho}. \end{aligned}$$

Then

$$\text{(3.15)} \qquad \hat{u}_1(\beta_1) = \hat{\rho}/\hat{n}^*$$

is an estimate of u_0, while

$$\text{(3.16)} \qquad \hat{u}_2(\beta_1) = \left(\frac{m_0(\beta_1)}{\hat{n}^*} - 0.5\right)/0.3989$$

is another estimate of u_0. Since β_1 is unknown, we vary the trial value β, computing $\hat{u}_1(\beta)$ and $\hat{u}_2(\beta)$.

As β increases around β_1, m_0 tends to increase rapidly, m_1 decreases less rapidly and m_2 is pretty stable. Analysis indicates that $\hat{u}_2(\beta)$ tends to increase more rapidly than $\hat{u}_1(\beta)$ when $\sigma\tau$ is small and $(1-p)/p$ is not large. Thus, the two values will tend to cross at a rather clearly defined estimate $\hat{\beta}_1$ and β_1. The above statement may be a bit of an exaggeration, since the stochastic behavior of $\hat{u}_1(\beta)$ and $\hat{u}_2(\beta)$ may lead to several crossings in a small neighborhood of $\hat{\beta}_1$, although the asymptotic expected values of $\hat{u}_1(\beta)$ and $\hat{u}_2(\beta)$ intersect with sharply different slopes at β_1.

The estimation procedure described above is somewhat ad hoc, and its extension for the case of unknown σ is not perfectly clear. However, it can be replaced by a more formal maximum likelihood estimation (MLE) procedure based on m_0, m_1, and m_2 where, for a trial value β of β_1, $Z_i = Y_i - \beta X_{i1}$ is assumed to have, conditional on $Z < 2.6\sigma$, the approximate c.d.f.

$$F_1(z;\theta,n^*,\rho) = \frac{n^*\Phi(\frac{z-\theta}{\sigma}) + \rho\psi(\frac{z-\theta}{\sigma})}{n^*\Phi(2.6 - \frac{\theta}{\sigma}) + \rho\psi(2.6 - \frac{\theta}{\sigma})} . \tag{3.17}$$

Then β_1 would be estimated by $\hat{\beta}_1$ where

$$(\beta - \hat{\beta}_1)\,\bar{X}_1 = \hat{\theta} \tag{3.18}$$

and $\bar{X}_1$ is the average of the X_{i1} for which $X_{i1} > c$ and $Z_i < 2.6\sigma$. If $\beta - \hat{\beta}_1$ is large, this procedure can be iterated with the first $\hat{\beta}_1$ used as the new trial value of β.

This MLE method can be extended by using the frequencies $m_j(\beta)$ for more than the three levels $z/\sigma = 0$, 1, and 2.6. Indeed, for the case of unknown σ, the use of MLE naturally suggests the application of at least four such levels.

The choice of 0, 1 and 2.6 as the coefficients of σ in (3.10) to (3.12) was somewhat arbitrarily made. Asymptotic analysis based on the model should provide optimal choices of the coefficients in terms of p, σ, and τ. A good choice for the ad hoc method tends to make $\hat{u}_1$ and $\hat{u}_2$ have maximally different slopes at β_1. For the MLE method one would maximize the Fisher Information.

Other sample properties have potential value. The function G was introduced to compare the conditional expectation of $Z = Y - \beta_1X_1$ given $z < c\sigma$ with the corresponding sample mean. Some such properties may be useful in developing a generalization of this ad hoc method for the case where σ is not assumed known.

In the example considered in Section 2, an analysis of the distribution of V for $X_1 > 0.2$ suggests that the mixture model we have used in Section 3 fits moderately, but not very, well. The lack of fit does not seem to cause much difficulty and the procedure based on this model seems moderately robust. On the other hand, it is easy to construct an alternative model that may fit a little better where the distribution of Z is a mixture of 0 and of an exponential distribution convolved with a normal. The asymptotic properties of the MLE based on the use of one of these mixture models when a possibly different model applies can be derived by use of the methods of Huber (1966).

REFERENCES

Huber, P. (1966). The behavior of maximum likelihood estimates under nonstandard conditions. _Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability_ Vol. 1, Univ. of Cal. Press.

Rutan, S.C. and Brown, S.D. (1984). Adaptive Kalman filtering used to compensate for model errors in multicomponent methods, _Analytica Chimica Acta_, 160 99-119.