INTERPRETATION OF REGRESSIONS WITH MULTIPLE PROXIES

Darren Lubotsky and Martin Wittenberg*

Abstract—Multiple proxy variables are typically available for an unobserved explanatory variable in a regression. We provide a procedure by which the coefficient of interest can be estimated from a regression in which all the proxies are included simultaneously. This estimator is superior in large samples to the common practice of creating a summary measure of the proxy variables. We examine the relationship between parents' income and children's reading test scores in the United States, and between parents' assets and children's school enrollment in India, and demonstrate that the reduction in attenuation bias from a better use of proxy variables can be significant.

I. Introduction

MEASUREMENT error in an independent variable in a regression model and the resulting attenuation bias in the coefficient estimate is one of the most well-known problems in empirical work. Although a great deal of attention has been paid to the effects of a single mismeasured independent variable, much less is known about the analytics and empirical strategies when more than one measure or proxy of the variable is available. It stands to reason that when more information is available, the problem of attenuation should be reduced. The question is how to do this in the best possible way.

In this paper we show that the way in which additional measures are currently incorporated in applied work is generally ad hoc and hardly ever optimal. Most commonly, researchers enter in a regression a single summary measure created from their set of proxy variables. We propose a superior method in which the proxies are entered separately in the regression and then a summary measure of their effect is created by combining their coefficients. To motivate this procedure, consider the following common empirical applications.

A. Permanent Income and Intergenerational Mobility

One example where the addition of more information seems to make a clear-cut difference is in the estimation of the effect of parents' permanent income on the education, health, and subsequent earnings of their children. The stronger the association between parental permanent income and child outcomes, the greater the degree to which income inequality is transmitted across generations. Estimating the precise magnitude of this association has proved difficult because permanent income is not observed. Instead, observed income in any year includes transitory components representing luck, measurement error, and other unanticipated shocks to income. A regression of children's outcomes on parents' observed income will understate the relationship between parental permanent income and child outcomes, understating the degree to which economic success or failure is transmitted across generations. To address this estimation problem, Solon (1992), Zimmerman (1992), and more recently Mazumder (2001) average parents' income over several years to arrive at a more precise measure of permanent income, and they show that the estimated regression coefficient increases markedly as more years of income are included in the average, suggesting that the problem of attenuation is reduced.

One potential problem with an average of annual income as a measure of permanent income is it assumes that incomes earned at different points in the life cycle are equally good measures of permanent income. If income earned earlier in life is a relatively noisier measure of permanent income, then it would seem a weighted average, with less weight given to income earned in earlier years, might do even better than a simple average. We will show below how such weights can be computed and that a better use of annual earnings data increases the estimated effect of parents' permanent income on children's reading test scores by over 30%, compared to an estimate derived from a simple average of annual earnings.

B. The Effect of Wealth on School Enrollment When Wealth is Not Observed

A more problematic case occurs when the variable of interest is simply not measured at all. For instance, the Demographic and Health Surveys are large household data sets with nearly identical questionnaires in over 40 developing countries, but they contain no information on respondents' income or wealth. To study many interesting questions about the determinants of health, educational attainment, or other household decisions requires income or wealth to be proxied by a variety of asset variables, such as whether or not the family owns a car or television, and the source of its home drinking water. Filmer and Pritchett (2001) suggest that the factor that accounts for the largest fraction of the variance in ownership across the assets is likely to be wealth and thus the first principal component of 21 such asset variables is a natural measure of household wealth. One problem with this procedure is that if ownership of each of the assets is a function of wealth and also of tastes or other characteristics of the household, the first

Received for publication April 16, 2003. Revision accepted for publication August 9, 2005.

* University of Illinois at Urbana-Champaign and University of Cape Town, respectively.

We thank two anonymous referees as well as Christopher Bollinger, Anne Case, Angus Deaton, Bo Honoré, Roger Koenker, Hashem Pesaran, Elie Tamer and seminar participants at the African Econometrics Conference, Michigan State University, the North American Summer Meeting of the Econometric Society, Princeton University, the Southern Economics Association annual meeting, and the University of Illinois for their comments and suggestions. We are naturally responsible for any errors. This project began while both authors were affiliated with the Research Program in Development Studies at Princeton University. We acknowledge financial support from the John D. and Catherine T. MacArthur Foundation through their Network on Poverty and Inequality in a Broader Perspective.

principal component will extract part of both wealth and tastes. There is no reason to believe that this composite will maximize the predictive power of the asset variables. We show below that a considerably stronger signal can be extracted, leading to an almost doubling of the regression effect of wealth on the probability of school enrollment in India, compared to that estimated by Filmer and Pritchett.

C. *Other Approaches to Utilizing the Information in Multiple Proxies*

The two examples above show differing approaches to the question of combining the information from different variables. These are not the only ways of trying to extract a stronger signal from various noisy sources of information, however. Glaeser et al. (2000), for instance, create an index of trust by standardizing (subtracting the mean and dividing by the standard deviation) responses to several survey questions and then adding them up. Mauro (1995) uses indexes of political and labor stability, "red tape," corruption, terrorism, and several other outcomes compiled by Business International, a private consulting firm, to measure institutional efficiency and corruption. Because he believes many of these indexes measure the same underlying phenomena, he averages the indexes together and uses the average as a regressor in models of growth and investment across countries. Herrnstein and Murray (1994) construct a measure of family socioeconomic status by averaging standardized values of parents' education, Duncan occupational scores, and family income. Similar examples are common in many fields of applied research.

Several considerations underlie the authors' strategies to summarize the proxies in a single, new variable. Firstly, the measurement error problem may be reduced by taking some linear combination of the proxies. As Mauro (1995) notes:

> Part of the rationale for aggregating the indices into composite subindices is that there may be measurement error in each individual index, and averaging the individual indices may yield a better estimate of the determinants of investment and growth.

Secondly, researchers may be worried about multicollinearity. If the different proxies are in fact all measuring the same underlying phenomenon, then there is only one structural coefficient to be estimated. Putting multiple proxies in the regression may likely result in many insignificant individual coefficients.

Thirdly, the coefficient on a single summary of the proxies may be more readily interpretable. To continue the example from the Demographic and Health Surveys, it is not clear how to infer the effect of household wealth on children's education from the coefficients on variables indicating ownership of a television or the availability of running water in the home. In other cases, a researcher may enter a summary measure of the proxies to compare its coefficient with the coefficient on another variable in the model or with estimates from different models.

We propose a new estimation method and in doing so show that these concerns are incorrect, but incorrect in interesting ways. To use an index or summary measure created from the proxies that extracts the largest possible signal requires the researcher to know the relative degree of noise contained in each proxy variable, as well as the correlation in noise across variables. Without knowledge of these magnitudes, it is impossible to create the optimal summary measure from the proxy variables. Using the proxies simultaneously in a multiple regression delivers this information as part of the regression coefficients, and the researcher then simply combines the coefficients in a known way to obtain the estimate of the effect of the latent factor.

Our procedure is best thought of as a method to *interpret* the coefficients in a regression under the null hypothesis that the variables are all generated by a common latent factor. A virtue of the procedure is its transparency. If the null hypothesis is not true, then the regression is not invalid, only some of the inferences that can be drawn from it. By contrast, data manipulations done before the regression can obviously not be undone by a skeptical reader.

The plan of our discussion is as follows. In the following section we will introduce the basic problem we wish to investigate and the related literature. We do so by means of a simplified case with two proxy variables for a single unobserved factor. The main theoretical results are in section III. We present our main theorem, in which we prove the optimality of putting all the proxies into the regression, and discuss the interpretation of the result. In section IV we consider how our findings are affected by the introduction of correctly measured covariates. Then in section V we return to the examples described above and show that our estimator performs considerably better than the approaches previously used in the literature. We conclude by pointing to a number of open questions; an appendix contains the proofs of our main results.

II. The Basic Problem

The circumstances that we wish to investigate can be highlighted by means of the following equations:

$$y = \beta x^* + \varepsilon, \tag{1a}$$

$$x_1 = x^* + u_1, \tag{1b}$$

$$x_2 = \rho_2 x^* + u_2, \tag{1c}$$

where β, relating y and x^* in equation (1a), is the parameter of interest. We assume that x^* is unobserved, but that we have the two observed proxies x_1 and x_2. We further assume that x^* is uncorrelated with ε, and u_1 and u_2 are uncorrelated with x^* and ε. That is, the proxy variables do not have independent effects on y. These assumptions are routinely imposed when researchers assume they have classical

measurement error. Finally, we allow the covariance between u_1 and u_2, denoted by σ_{12}, to be unrestricted.

If we regress y on the first proxy, we have the well-known case of classical measurement error with the attendant attenuation bias. The OLS estimator b of β will converge asymptotically to

$$b = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_1^2}, \tag{2}$$

where $\sigma_x{}^2 = \text{var}(x^*)$ and $\sigma_1{}^2 = \text{var}(u_1)$. The parameter β is not identified. As Aigner et al. (1984) note, we may be able to identify β from higher-order moments if the latent variable x^* is not normally distributed.

In order to identify the parameter, in general we need one more restriction. If we set $\text{cov}(u_1, u_2) = 0$, then we could use x_2 as an instrument for x_1 in the regression of y on x_1. A different type of restriction is if we are able to measure σ_1^2. This is possible, for instance, if the second proxy variable is actually a repeat measurement, carried out for a subset of observations under controlled conditions. In this case we can obviously also retrieve σ_x^2 and then correct the OLS estimates. This is the *errors-in-variables* estimator (see Fuller, 1987).

We might consider whether the relationships between the proxies allow us to identify the common *factor* x^*. This is the domain of factor analysis (see Aigner et al., 1984). In general this requires us to make some additional assumptions. We need to impose orthogonality between the error variances u_1 and u_2, and also adopt a normalization on the coefficients. The factors so isolated are only identified up to multiplication by an orthogonal matrix.

Principal-components analysis achieves a unique decomposition, but does so by the expedient of identifying the common factor with the linear combination of proxies that maximizes the explained variance. It is not clear why this concept should correspond to the structural relationships underlying equations (1a)–(1c). Indeed, if the assumption of orthogonality between the error components fails, then this procedure is guaranteed to produce a composite of the factor x^* and the commonality in the errors.

Other identification strategies involve adding equations or specifying the process which generates the latent variable. In the MIMIC (multiple indicators, multiple causes) model, for example (see inter alia Aigner et al., 1984; Goldberger, 1972; Jöreskog and Goldberger, 1975), it is assumed that at least one more relationship is available between an indicator variable and the latent variable, parallel to that in equation (1a). The latent variable itself is written as a function of a set of observable variables, that is, equations (1b) and (1c) are replaced by

$$x^* = \alpha_1 x_1 + \alpha_2 x_2 + u_3,$$

and the proportionality relationships between the different equations are exploited to achieve identification. A generalization of this approach is the LISREL model (see Bollen, 1989). In this model the structural parameters are identified by cross-equation restrictions.

We assume that these sorts of strategies are not available for the cases under consideration. In particular, we assume that the researcher is not willing to make additional assumptions beyond those already given for equations (1a) through (1c), and the empirical problem is how best to use the proxy variables to minimize the attenuation bias, if not eliminate it. The issue therefore is how to make the best of a bad situation. Throughout, however, we restrict attention to models that are linear in the parameters.

Leamer (1983, pp. 314–315) has a discussion of "proxy searches" in which he addresses precisely this issue. His discussion is, however, exclusively about how to decide which one of the two (or more) proxies to include in the regression. His advice is to pick the variable which yields a high R^2 and which has a low variance. He does not consider whether one could do better by combining the information from the proxies.

In order to hone our intuition, let us consider the system in equations (1a)–(1c) with $\rho_2 = 1$. The covariance matrix of x_1 and x_2 is given by

$$\sum_{XX} = \begin{bmatrix} \sigma_x^2 + \sigma_1^2 & \sigma_x^2 + \sigma_{12} \\ \sigma_x^2 + \sigma_{12} & \sigma_x^2 + \sigma_2^2 \end{bmatrix},$$

and the covariance matrix of u_1 and u_2 by

$$\sum_{UU} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

with $\sigma_{12} \neq 0$. By our assumptions $\text{cov}(y_1, x_1) = \text{cov}(y_1, x_2) = \beta\sigma_x^2$, and hence the coefficients estimated from regressing y on proxy 1 or proxy 2 are given asymptotically, respectively, by

$$b^1 = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_1^2} \quad \text{and} \quad b^2 = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_2^2}.$$

Because the denominator is just the variance of the proxy variable, it is clear that the proxy with the smaller variance will give the least biased results.

What were to happen if we were to take a simple average of the two proxies? In this case

$$\frac{1}{2}(x_1 + x_2) = \bar{x} = x^* + \bar{u}$$

with $\text{var}(\bar{u}) = \frac{1}{4}(\sigma_1^2 + \sigma_2^2 + 2\sigma_{12})$. There clearly is no necessity that this be smaller than the minimum of σ_1^2 and σ_2^2. In particular, if one proxy is a good one and the other much worse, simply averaging them is unlikely to be the optimal strategy. Other linear combinations of the variables are likely to get a much better reduction in the error variance. Indeed, let $u_0 = \delta_1 u_1 + \delta_2 u_2$ be a linear combination such that $\delta_1 + \delta_2 = 1$; then it is straightforward to show that the choice of δ_1 that will minimize the variance of the weighted average is given by $\delta_1 = (\sigma_2^2 - \sigma_{12})/(\sigma_1^2 - 2\sigma_{12} + \sigma_2^2)$. The variance of the

linear combination in this case will be $(\sigma_2^2\sigma_1^2 - \sigma_{12}^2)/(\sigma_1^2 - 2\sigma_{12} + \sigma_2^2)$, so that the estimate of β is given asymptotically by

$$b^* = \beta \frac{\sigma_x^2}{\sigma_x^2 + \dfrac{\sigma_2^2\sigma_1^2 - \sigma_{12}^2}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}}.$$

No other linear combination of x_1 and x_2 will produce an estimate of β with less attenuation bias. Unfortunately, we do not know the variances and covariance of u_1 and u_2, and as it stands cannot compute this optimally weighted average of the proxies.

What happens if we run the regression of y on both proxies? The multiple regression coefficients will be given asymptotically by $\Sigma_{XX}^{-1}\Sigma_{Xy}$, where

$$\sum_{Xy} = \begin{bmatrix} \beta\sigma_x^2 \\ \beta\sigma_x^2 \end{bmatrix}.$$

It is straightforward to show that the coefficients on x_1 and x_2 are

$$b_1 = \beta \frac{\sigma_x^2(\sigma_2^2 - \sigma_{12})}{\sigma_x^2\sigma_1^2 + \sigma_x^2\sigma_2^2 - 2\sigma_x^2\sigma_{12} + \sigma_1^2\sigma_2^2 - \sigma_{12}^2}, \tag{3a}$$

$$b_2 = \beta \frac{\sigma_x^2(\sigma_1^2 - \sigma_{12})}{\sigma_x^2\sigma_1^2 + \sigma_x^2\sigma_2^2 - 2\sigma_x^2\sigma_{12} + \sigma_1^2\sigma_2^2 - \sigma_{12}^2}. \tag{3b}$$

This does not look very promising, but note that

$$\begin{aligned} b_1 + b_2 &= \beta \frac{\sigma_x^2(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})}{\sigma_x^2\sigma_1^2 + \sigma_x^2\sigma_2^2 - 2\sigma_x^2\sigma_{12} + \sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ &= \beta \frac{\sigma_x^2}{\sigma_x^2 + \dfrac{\sigma_2^2\sigma_1^2 - \sigma_{12}^2}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}} \\ &= b^*, \end{aligned} \tag{4}$$

so that adding up the coefficients of the two variables yields an estimate that is precisely equal to the coefficient obtained on the optimally weighted combination of the proxies. What is even more remarkable is that we did not need to know anything about the relative magnitudes of error variances and covariances in order to achieve this result—the regression accomplished this by itself.

We will show in the next section that this result holds true more generally—the attenuation bias can be minimized when all the proxies are entered separately in a multiple regression. Without additional assumptions or data, one cannot in general compute an estimate of β that is guaranteed to be less biased than b^*. We need to proceed with some care, however, in the situation where $\rho_2 \neq 1$. In this case the two proxy variables are not equally correlated with the latent factor. In general, then, we will want to take a weighted sum of the regression coefficients.

III. The General Case

This section considers the more complicated situation in which there are an arbitrary number of available proxies for x^* and the proxies may have different degrees of correlation with x^*. We begin by demonstrating that although β is generally not identified, attenuation bias is minimized by including the proxy variables separately in a regression and then constructing a weighted sum of the coefficients on each proxy variable, where an estimate of ρ_j is the weight given to coefficient j. We discuss the interpretation of regressions with multiple proxies, relate the new estimator to index construction, and conclude the section with a discussion of an assessment of an alternative instrumental variables estimation strategy.

We now assume that we have k proxies, with

$$x_j = \rho_j x^* + u_j$$

for $j = 1, \ldots, k$. We continue to assume that cov $(u_j, \varepsilon) = 0$, cov $(u_j, x^*) = 0$ for all j, but now assume that the covariance matrix of the u_j's is unrestricted; that is,

$$\sum_{UU} = E(U'U) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_k^2 \end{bmatrix},$$

where $U = [u_1\ u_2 \ldots u_k]$.

A. *Identification of* ρ *and Nonidentification of* β

We note that as it stands the coefficients ρ_j are not identified. The same observations of y and the x_js would be consistent with appropriately rescaled vectors ρ and β. Consequently we adopt the normalization [already used in equation (1b)] that $\rho_1 = 1$. This amounts to fixing the scale of the latent variable x^* in terms of the observable x_1. Other normalizations are possible, such as $\|\rho\| = 1$. None of our results depend on the precise normalization adopted. Whichever scale is adopted, our procedure will lead to the smallest attenuation bias in the estimates.

The scale is important, however, when comparing the magnitude of estimates across models. In some applications there may be a natural way to select the scale of the latent variable. For example, in our permanent-income example below, we use income earned when the child's mother is 22 and 23 years old as a scale for permanent income. Models that have appeared in the literature tend to scale variables implicitly by the average of income over the lifetime, which is a larger scale than our choice. Consequently, we need to rescale our estimates to make a valid comparison with those that have appeared in the literature. In other applications there may not be a natural scale. For example, in our example of household wealth and school enrollment in India, we choose the number of rooms in a house as a scale for wealth.

The available information is contained in the covariance matrix

$$\sum_{ZZ} = E(Z'Z) = \begin{bmatrix} \beta^2\sigma_x^2 + \sigma_\varepsilon^2 & \beta\sigma_x^2 & \beta\rho_2\sigma_x^2 & \cdots & \beta\rho_k\sigma_x^2 \\ \beta\sigma_x^2 & \sigma_x^2 + \sigma_1^2 & \rho_2\sigma_x^2 + \sigma_{12} & \cdots & \rho_k\sigma_x^2 + \sigma_{1k} \\ \beta\rho_2\sigma_x^2 & \rho_2\sigma_x^2 + \sigma_{12} & \rho_2^2\sigma_x^2 + \sigma_2^2 & \cdots & \rho_2\rho_k\sigma_x^2 + \sigma_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta\rho_k\sigma_x^2 & \rho_k\sigma_x^2 + \sigma_{1k} & \rho_2\rho_k\sigma_x^2 + \sigma_{2k} & \cdots & \rho_k^2\sigma_x^2 + \sigma_k^2 \end{bmatrix},$$

where $Z = [y\ x_1\ x_2 \ldots x_k]$.

There are altogether $k(k+1)/2$ unknown parameters in Σ_{UU}, $k - 1$ parameters in ρ, and the parameters β, σ_x^2, and σ_ε^2. Because there are altogether $(k + 1)(k + 2)/2 + 1$ unknown parameters in Σ_{ZZ}, but only $(k + 1)(k + 2)/2$ pieces of observable information, we are short one restriction in order to identify the parameter β.

Observe, however, that the vector ρ is identified from the covariances between the dependent variable y and the proxies:

$$\rho_j = \frac{\text{cov}\ (y,x_j)}{\text{cov}\ (y,x_1)} \tag{5}$$

As noted above, there are several ways in which we could achieve identification of β: any restriction on the covariance matrix of $X = [x_1\ x_2 \ldots x_k]$ will do so in principle. Zero restrictions on any of σ_{1j} would allow us to use x_j as an instrument for x_1. More generally, a zero restriction on σ_{jh} would allow us to use x_h as an instrument for x_j, but the resulting estimate would need to be rescaled to take account of the fact that x_j is not on the same scale as the latent variable x^*. Because we have an estimator for ρ_j, this is easily achieved. We have

$$\beta = \frac{\text{cov}\ (y,x_h)}{\text{cov}\ (x_j,x_h)} \frac{\text{cov}\ (y,x_j)}{\text{cov}\ (y,x_1)}.$$

The first factor is the instrumental variables estimator; the second is the GMM estimator of ρ_j.

If we know the magnitudes of σ_x^2 or of any of the error variances or covariances, we could construct a generalization of the errors-in-variables estimator. As in the previous section, however, we will assume that we do not have any plausible restrictions. In this case the issue is how to optimally use the information contained in the proxies in order to minimize the attenuation bias.

B. *Minimizing Attenuation Bias*

We begin by introducing some additional notation. Let

$$x^\delta = X\delta$$

be any linear combination of the proxy variables, where $X = [x_1\ x_2 \ldots x_k]$. By construction

$$X = x^*\rho' + U,$$

where $\rho' = [1\ \rho_2 \ldots \rho_k]$. It follows that

$$x^\delta = x^*\rho'\delta + U\delta, \tag{6}$$

and hence the structural model in equation (1a) can be written as

$$y = \frac{\beta}{\rho'\delta}x^\delta + \varepsilon - \frac{\beta}{\rho'\delta}U\delta.$$

Our two-variable example above assumed that $\rho_1 = \rho_2 = 1$ and $\delta_1 + \delta_2 = 1$, which implies $\rho'\delta = 1$. In general there is obviously no reason why one might expect that $\rho'\delta = 1$. One could interpret a value of $\rho'\delta = \kappa \neq 1$ as a rescaling of the latent variable x^* or equivalently of the parameter β. In the presence of attenuation bias this rescaling could exacerbate or reduce the attenuation bias. In the absence of additional information, however, the direction of the impact is unknowable.

Our main result—contained in the theorem below—is that simply including the proxy variables separately in a regression and then aggregating their coefficients produces the estimate with the minimum attenuation bias among the class of composite proxy variables in which $\rho'\delta = 1$.[1] Other combinations of proxy variables in which $\rho'\delta = \kappa \neq 1$ may do better or worse, but it is not possible to characterize the bias without additional assumptions. However, we can map all these composites into the set where $\rho'\delta = 1$. In this case the direction of bias becomes known and the condition of our theorem applies. Formally, the generalization of the discovery we made in the two proxy case is contained in the following theorem:

Theorem 1. Let $\hat{\beta}$ be the OLS estimate in the regression of y on x^δ, let b be the OLS estimate in the regression of y on X, and $\hat{\rho}$ be the GMM estimate of ρ. Then

$$|\text{plim}\ \hat{\beta}(\rho'\delta)| \leq |\text{plim}\ \hat{\rho}'b| \leq |\beta|,$$

where

$$\begin{aligned} \text{plim}\ \hat{\beta}(\rho'\delta) &= \beta\left(1 - \frac{\delta'\Sigma_{UU}\,\delta}{\sigma_x^2(\delta'\rho)^2 + \delta'\Sigma_{UU}\,\delta}\right), \\ \text{plim}\ \hat{\rho}'b &= \beta\left(1 - \frac{\left|\sum_{UU}\right|}{\left|\sum_{XX}\right|}\right) \\ &= \beta\left(1 - \frac{1}{\sigma_x^2\rho'\sum_{UU}^{-1}\rho + 1}\right), \end{aligned} \tag{7}$$

[1] We thank one of our referees for sharpening our discussion on this point.

and Σ_{XX} is the covariance matrix of X. For every $\delta \neq 0$ we have

$$\frac{1}{\sigma_x^2 \rho' \sum_{UU}^{-1} \rho + 1} \leq \frac{\delta' \Sigma_{UU} \delta}{\sigma_x^2 (\delta' \rho)^2 + \delta' \Sigma_{UU} \delta}.$$

Equality holds only if

$$\delta = c \sum_{UU}^{-1} \rho$$

for some $c \neq 0$.

The proof is contained in the appendix.

Three points follow from this theorem. First, if $\rho'\delta = 1$, the researcher can do no better than including the proxies simultaneously in the regression and then aggregating their coefficients. Second, if $\rho'\delta = \kappa \neq 1$ and no additional information is available, one will get unpredictable results: The point estimate $\hat{\beta}$ may be above or below the true value of β, but one can say little more than that. However, one can always rescale x^δ by dividing it by $\rho'\delta$ (this rescaling is feasible because ρ can be consistently estimated and δ is chosen by the researcher). Inspection of equation (6) reveals that this rescaling turns an ad hoc combination of the proxy variables into a case of classical measurement error. Because dividing x^δ by $\rho'\delta$ is equivalent to multiplying $\hat{\beta}$ by $\rho'\delta$, theorem 1 says that rescaling an ad hoc combination of proxies into a case of classical measurement error will lead to a coefficient estimate that contains a larger attenuation bias than $\hat{\rho}'b$.

Third, there is in principle always some rescaling that eliminates the attenuation bias. For instance in our two-variable example above with $\rho_1 = \rho_2 = 1$, if $\delta_1 = \sigma_x^2/(\sigma_x^2 + \sigma_1^2)$ and $\delta_2 = 0$, then a regression of y on $\delta_1 x_1$ will produce an unbiased estimate of β. This rescaling of the first proxy variable is equivalent to the errors-in-variables rescaling of the OLS estimator. This, however, requires additional information, namely the value of σ_1^2 or σ_x^2.

The theorem therefore covers all of the procedures outlined in the introduction—averaging, standardizing, and then adding—and construction of the first principal component. In any particular empirical application, a researcher using one of these traditional methods for dealing with multiple proxy variables may compute an unbiased estimate, or an estimate that is less biased than $\hat{\rho}'b$. Unfortunately, though, there is no way to know whether the estimate is upward or downward biased relative to β, or whether it is, in fact, less biased than $\hat{\rho}'b$. If, however, we rescale these procedures so that the direction of bias becomes known, they will have an attenuation bias that is strictly larger than the attenuation bias from our procedure.

We reiterate that the appropriate way of aggregating up the coefficients on the k proxy variables in the multiple regression is given by

$$b^\rho = \hat{\rho}'b = \sum_{j=1}^{k} \frac{\text{cov}(y,x_j)}{\text{cov}(y,x_1)} b_j, \tag{8}$$

where b_j is the coefficient on x_j in the multiple regression.

Another way to understand this result is to note that the formula in equation (7) provides the natural generalization of the errors-in-variables formula in equation (2), with the *generalized variances* $|\Sigma_{UU}|$ and $|\Sigma_{XX}|$ replacing σ_1^2 and σ_x^2 respectively (see Dhrymes, 1974, p. 56). As in the formula for the bias resulting from a single mismeasured independent variable, the size of the attenuation bias in the multiple-proxy-variable case depends on the ratio of the variance in the error components to the total variance in the proxy variables.

C. Regression Inference and Index Construction

Our procedure obviously depends on the validity of the underlying assumptions. If cov $(u_j, \varepsilon) \neq 0$ and the proxies belong in the main regression [equation (1a)], then clearly the process of aggregating up the coefficients will not correspond to any parameter of interest. Nevertheless, the procedure is more robust to departures from the validity of the underlying assumptions than will be index construction prior to estimating the regression.

An estimate of ρ can be reported with the regression results, and the reader can then assess the plausibility of the common-factor assumption. For example, if the latent variable is wealth, it would be strange if the number of rooms in one's house did not load strongly onto it. Furthermore, given ρ, it is possible to provide different estimates of β, depending on whether particular proxies are viewed as having independent effects or not. The transparency of this procedure is attractive.

Strictly speaking, our procedure should be viewed as an *interpretation* procedure, rather than an *estimation* procedure. By aggregating the k different regression coefficients, our procedure gives a way to interpret how changes in the underlying unobserved variable x^* affect the dependent variable y. The coefficients on the proxies themselves have the less straightforward interpretation of the effect of a unit change in the proxy holding all other proxies constant.

Another way in which we can interpret the procedure is as a particular way of constructing a composite index from the separate proxies. Indeed, as theorem 1 shows, there will always be one linear combination of the variables that will provide exactly the same coefficient as the estimator in equation (8). The multiple regression can therefore be viewed as implicitly constructing an index from the separate proxies. Our procedure provides the coefficient on this index. Indeed, we can make this implicit index explicit:

$$x^\rho = \frac{1}{b^\rho} \sum_{j=1}^{k} x_j b_j,$$

where b_j is the j^{th} regression coefficient. By construction this index is on the same scale as x_1 and will reproduce b^ρ as the coefficient in the regression.[2]

This index is the common factor in the proxies that best explains y. In a regression with a different dependent variable, a different index would be selected by the procedure. One should therefore be cautious in strictly identifying the index with the underlying latent variable. At the same time a virtue of the index is that it allows us to do various checks on the plausibility of the procedure. For example, the correlation structure between an index representing wealth and other variables can be compared with a similar correlation structure estimated from other data utilizing observed wealth.

If we view the regression as a procedure for implicitly constructing an index, then the individual regression coefficients have the interpretation as weights. From the final equality condition in theorem 1 it is clear that this reweighting must work so that the weight is proportional to the correlation with x and (in a sense) inversely proportional to the error variance. The multiple regression procedure must therefore parcel out the overall regression coefficient b^ρ proportional to ρ and inversely proportional to the error variance. We can show this somewhat more precisely.

Proposition 2. Let b_i be the i^{th} regression coefficient in the multiple regression of y on X, that is,

$$b_i = e_i'(X'X)^{-1}X'y,$$

where e_i is the unit vector with 1 in the i^{th} position. Then

$$\text{plim } b_i = \beta\frac{\sigma_x^2\left|\sum_{UU}^{\rho(i)}\right|}{\left|\sum_{XX}\right|}, \tag{9}$$

where $\Sigma_{UU}^{\rho(i)}$ is the matrix obtained by deleting row i of Σ_{UU} and replacing it with the vector ρ'.

The proof is contained in the appendix.

In the special case where Σ_{UU} is the diagonal matrix, it follows that

$$\text{plim}\frac{b_i}{b_j} = \frac{\rho_i\prod_{k\neq i}\sigma_k^2}{\rho_j\prod_{k\neq j}\sigma_k^2} = \frac{\rho_i\sigma_j^2}{\rho_j\sigma_i^2}.$$

Several additional points follow from this result. Firstly, if $\beta = 0$, then plim $b_i = 0$. This means the hypothesis that $\beta = 0$ is testable as a joint hypothesis that the coefficients on all the proxies are jointly 0. Because $|\text{plim } \hat{\rho}'b| \leq |\beta|$, one could also test whether $\hat{\rho}'b$ is statistically different from 0. Unfortunately, without knowing the degree of attenuation bias in $\hat{\rho}'b$, the true significance level or the power of this latter test cannot be determined. It bears pointing out that even if the researcher is not concerned with the magnitude of β and simply wants to test for statistical significance, it is still preferable to obtain the least attenuated estimate as possible to maximize the chance of rejecting a false null hypothesis.

Secondly, it follows from our proof of theorem 1 that

$$\left|\sum_{XX}\right| = \left|\sum_{UU}\right| + \sum_{i=1}^{k}\rho_i\sigma_x^2\left|\sum_{UU}^{\rho(i)}\right|$$

There are therefore $k + 1$ terms in the denominator of equation (9). If the proxies are all of similar quality (that is, if the ρ's and error variances are not vastly dissimilar), then the individual coefficients should be of the order of β/k. As more proxies are added, the individual coefficients should tend to 0. It is this feature that possibly accounts for researchers' apprehension in adding multiple noisy measures of the same variable into a regression. It should be clear, however, that this is not the appropriate metric in which to think about the size of the coefficient. It is not the individual contributions that matter, but the aggregate one.

D. Biased Instrumental Variables Estimation

The results above indicate that including proxy variables simultaneously in a regression is superior to the ad hoc index construction approaches seen in the literature. Nevertheless, it is not clear that it is the best approach possible. For instance, if the error components in the proxies are mutually independent, then instrumental variables will deliver an unbiased estimate of the structural parameter. One might speculate, therefore, that biased IV estimation in cases where the errors in the proxies are only weakly correlated might still do better than our recommended approach.

If two proxy variables are available—x_1 and x_2 defined as above—then biased IV estimation may indeed have a smaller asymptotic bias than our proposed estimator. Whether the bias is larger or smaller in a particular situation depends on the magnitude of the variances and covariances of the error components u_1 and u_2. If $k > 2$ proxy variables are available, then the researchers must choose the optimal linear combination of $k - 1$ proxies for which the error component is least correlated with u_1. The conventional two-stage least squares estimate will definitely fare badly in this regard, because it will seek to explain both the part of x_1 which is correlated with x^* and also the error term u_1. Without making additional assumptions, it is impossible to construct the optimal instrument or ascertain whether the resulting bias is larger or smaller than the bias from our proposed estimator. Furthermore, the direction of the IV bias is indeterminate. Consequently, reporting both the estimates derived from a least squares regression with all

[2] We have $\delta = (1/b^\rho)\, b$ with $b = (X'X)^{-1}X'y$. Consequently $\delta'\rho = 1$ (because $b^\rho = b'\rho$) and $(\delta' X' X \delta)^{-1}\delta'X'y = b^\rho$.

proxies included as well as any IV estimates is likely to be a preferred strategy.[3]

4. The Effect of Covariates

It is well known that the attenuation bias in the OLS coefficient on a mismeasured variable is increased when correctly measured variables are also included in the model, provided that these variables are not correlated with the measurement error (Griliches, 1986). Furthermore the bias is transmitted to the coefficients of the correctly measured variables, generally with the opposite sign. These results apply in our case as well, as is shown in the following proposition:

Proposition 3. If the true model is given by

$$y = x^*\beta + Z\gamma + \varepsilon,$$

where Z is a matrix of correctly measured covariates, then the estimator $\hat{\rho}'b$ provides an estimate of β that has lower attenuation bias than any other linear combination of the proxies.

If Z is uncorrelated with any of the measurement error components u_j, the degree of attenuation bias is given by

$$\frac{|\sum_{UU}|}{|\sum_{M_ZXM_ZX}|} \geq \frac{|\sum_{UU}|}{|\sum_{XX}|},$$

where $\Sigma_{M_ZXM_ZX}$ is the covariance matrix of the residuals when X is regressed on Z. That is, attenuation bias is increased in the presence of covariates. Furthermore, the estimate of γ will be biased.

If Z is correlated with any of the measurement error components, then the presence of covariates may increase or decrease the degree of attenuation bias, but the bias in γ will be exacerbated.

The proof of this result is based on repeated applications of the Frisch-Waugh-Lovell theorem (Davidson & MacKinnon, 1993, pp. 19ff). The details are contained in the appendix. Essentially, by regressing all variables on Z and obtaining the residuals, we get a system of equations which looks like the case considered in the previous section. Indeed, our original model should be thought of as the model obtained when all variables have been regressed on a constant, so that all variables are expressed as deviations from their mean. The only complicating factor when covariates are present is that the estimator of ρ may need to be adjusted if it is likely that the covariates are correlated with the measurement error term. In this case ρ is estimated using the residuals from a regression of the proxies and dependent variable on the covariates, rather than the actual variables.

In short, our conclusion that it is better to include the proxy variables separately in the regression and then aggregate the coefficients still holds when correctly measured covariates are present in the model. However, in this situation our procedure should not be taken as license to throw any and all possible proxy variables into the regression. It is particularly important to be concerned about the correlation between covariates and the measurement error component in the proxy variables. Adding proxies that absorb the effects of the covariates instead of proxying for the latent variable will be particularly damaging. An important tradeoff exists, therefore, in adding proxies that may add little information about the underlying unobserved variable, but affect the accuracy with which we measure the coefficients on correctly measured variables in the model. Ideally, the proxies should be correlated with x^*, and their measurement error components should be orthogonal to the other explanatory variables in the regression. [See Bollinger (2003) for further information on how the presence of proxy variables affects the estimation of correctly measured variables in a regression.]

V. Applications to Existing Research

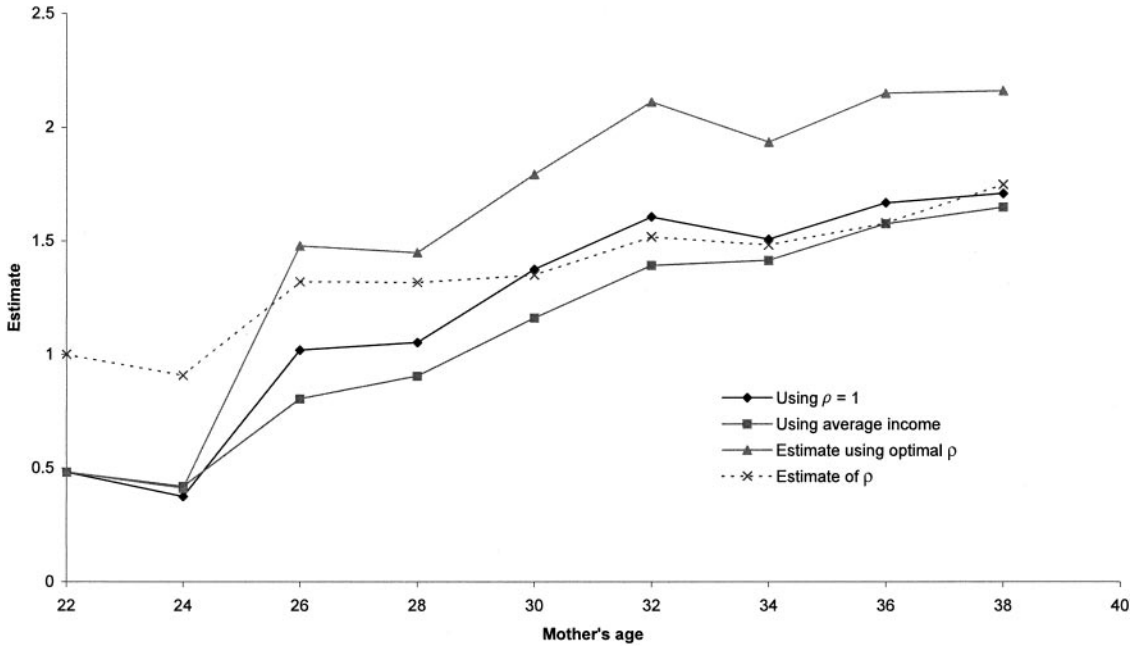
We illustrate the procedure with two empirical examples. Though we do not know the true data-generating process and cannot compare the alternative estimators against the true parameter values, the use of actual data allows us to assess whether the alternative estimators themselves produce qualitatively different results. In the first example we are interested in estimating the relationship between parents' permanent income and children's performance on a reading comprehension test. Permanent income is not observed, and we instead have panel data on annual family income. In the second example we use data on assets and housing conditions from the Demographic and Health Survey of India as proxies for household wealth in a model linking wealth and school attendance. Filmer and Pritchett (2001) use the first principal component of the asset variables as their measure of wealth.

Economists and other social scientists have long been interested in determining the effect of parental income on measures of child development and their eventual labor market earnings. A greater association between parent and child outcomes implies that a greater degree of economic success and failure is transmitted across generations. Empirical attempts to measure this association have recognized for some time that because observed parental income is a noisy measure of their permanent income, the association between observed income and child outcomes will understate the association between permanent income and child outcomes. This will lead to an understatement of the intergenerational correlation in economic success. To circumvent this problem, a general practice in the literature has been to average annual income over several years to arrive at a less noisy measure of parental income. See, for example, Blau (1999), Case, Lubotsky, and Paxson (2003), Mayer (1997), Solon (1992), and Zimmerman (1992).

Although past work has made clear that using more years of parental income leads to an increase in the measured association between parental income and child outcomes,

[3] Further details on biased instrumental variable estimation are contained in a previous version of the paper. See Lubotsky and Wittenberg (2003).

FIGURE 1.—THE EFFECT OF FAMILY INCOME ON CHILDREN'S READING COMPREHENSION SCORE



Note: Data are from the NLSY-Children, 1979–1998. All models also include controls for the log family size; the child's sex, age, and race; the mother's age and education; whether the mother's spouse is present, and if so, his age and education; year effects; and the mother's AFQT score. The estimates using the optimal p have been divided by the average of the ρ's in order to put them on the same scale as the other two estimators.

our estimation strategy can potentially do even better. A standard model of log permanent income specifies observed log income at age t as being a function of unobserved permanent income (y^p), life cycle effects, and transitory or luck components: $y_t = \rho_t y^p + u_t$, where ρ_t represents the age-earnings profile, capturing the idea that younger workers tend to earn below their level of permanent earnings; and u_t reflects deviations from age-adjusted permanent earnings, which may be serially correlated and heteroskedastic across time. Simply averaging y_t across time ignores life cycle variation in earnings, covariance between the measurement error components over time, and differences in the variance of the error component over time. Our procedure is designed specifically to take these factors into account and arrive at a less noisy measure of parental permanent income.

Using data from the National Longitudinal Survey of Youth (NLSY), we examine the relationship between family income and children's percentile score on the Peabody Individual Achievement Test in reading comprehension. The NLSY began in 1979 with a sample of 12,686 individuals aged 14 to 21. Interviews were conducted annually between 1979 and 1994, and have been conducted biennially since then. In 1986 a separate biennial survey of the children of the women from the 1979 cohort began (called the NLSY-Children). Missing data pose a difficulty for including annual incomes separately in the regression. Therefore, we work with two-year averages of family income taken when the mothers were between the ages of 22 and 39. Our sample uses data collected between 1979 and 1998 and contains 4,668 child-year observations of those aged 6 to 14 and who have nonmissing family income during this period. The model also includes controls for the log of family size; the child's sex, age, and race; the mother's age and education; whether the mother's spouse is present, and if so, his age and education; year effects; and the mother's age-adjusted AFQT score (a test of reading and math skills that was administered to the mother in 1980). We drop children who have missing data for any of these controls.

Figure 1 plots the results from different models of children's test scores. Following the common practice in the literature, our first measure of permanent income is the average log income over several periods. The line in the figure labeled "Using average income" indicates the coefficients on this term when it contains income when the mother was aged 22–23 to the age indicated on the x axis. The coefficient rises from 0.5 when only income when the mother is 22 and 23 is used, to 1.2 when income between ages 22 and 31 is used, and finally to 1.6 when income through age 39 is used. Clearly (as prior researchers have found), more income data can reduce attenuation bias.

Next we include family income from different periods separately in the regression and average the coefficients, first with an unweighted average and then weighted by the GMM estimate of ρ. The unweighted average of the income coefficients produces a total effect that is in some cases 27% larger than the effect of average income. This difference indicates the importance of heteroskedasticity and serial correlation in u_t, the transitory component in annual earnings.

The GMM estimates of ρ are given in the figure and show a steady rise over the life cycle, consistent with the idea that earlier income understates permanent income. We use these estimates to optimally weight the separate income coefficients, and this leads to a substantial rise in the relationship between children's test scores and permanent income.[4] Using family income when the mother is aged 22 to 39, the effect from using the optimally weighted coefficients is 2.2, compared to only 1.6 when income is averaged prior to the regression, an increase of 31%. Our optimal estimator implies that an increase in permanent income of 0.8 (which is approximately the standard deviation of log income among 30–31-year-olds) is associated with a 1.76-percentile-point rise in test scores. The key feature of this example is that income earned later in life is a better measure of permanent income, and our estimates incorporate this additional information better than does the simple average of annual incomes. This leads us to estimate an intergenerational elasticity that is less attenuated than what previous researchers have found.

Our second empirical application reexamines Filmer and Pritchett's (2001) estimate of the effect of household wealth on Indian children's propensity to be enrolled in school, using data from the 1992–1993 Demographic and Health Survey of India. The difficulty is that the data do not contain any information on income or wealth, but do contain many questions on asset holdings and dwelling quality. Filmer and Pritchett propose to use the first principal component of these asset variables as their measure of wealth.

We use data on 109,973 children aged 4 to 16 with nonmissing data for all variables. The dependent variable in the regression is an indicator that the child is enrolled in school. The asset variables are the number of rooms in the house, and indicators for whether the household has a refrigerator, clock or watch, sewing machine, VCR, radio, television, fan, bicycle, car, motorcycle, electric lighting, flush toilet or latrine, and livestock; whether the kitchen is in a separate room in the house; whether the primary cooking fuel is wood, cow dung, or coal; and whether the drinking and nondrinking water comes from a pump or an open source (as opposed to being piped into the home).

The first column of table 1 displays our estimates of ρ. Each coefficient ρ_j can be interpreted as a structural regression coefficient in the regression of the proxy on the latent variable (wealth), rescaled so that one unit of wealth would lead to one additional room. Looking at these coefficients, it is clear that an increase in wealth will have a large impact on the probability of owning a radio or a television, and lead to a decrease in the probability of using wood or cow dung as fuel. The relatively low increase in the probability of owning a car or a VCR is due to the fact that ownership of these assets is very rare in this population.

[4] We divide the estimates using the optimally weighted coefficients by the average of the ρs in order to make them comparable in scale to the previous two estimators.

The next six columns show results when all or some of the asset variables are entered separately into the regression. The model also controls for the child's sex and age, the head of the household's sex, age, and education, and the log family size. Nearly all of the asset variables are statistically significant, although some, such as refrigerator, car, and VCR ownership and using wood, dung, or coal as cooking fuel, have a different sign (the b_j's) than their raw correlation with school enrollment and with their value of ρ_j. One might be tempted to drop these variables from the model, thinking they are capturing something other than the effect of wealth on school enrollment. As illustrated in equation (3), a proxy that is highly correlated with another, better-measured proxy may well have a different sign than the true effect to be measured (β). Dropping the variables discards useful information and is thus counterproductive.

When all 18 asset variables are used, the estimated effect of the assets is 0.170.[5] We bootstrapped the estimation procedure and estimate a standard error of 0.003.[6] To see how sensitive the estimate is to using fewer proxies and to different groups of proxies, in the next five columns we break the 18 proxies into two groups of nine and then three groups of six. When nine are used, the effects are 0.136 and 0.132; when six are used the effects are 0.129, 0.105, and 0.116. The attenuation bias in the estimates clearly increases as fewer proxies are used. The estimates that utilize the same number of proxies are remarkably close to each other, suggesting the assumption of a single unobserved factor is plausible. All estimates are statistically different from zero.

The last column of table 1 displays the scoring vector used to weight the asset variables for the first principal component. These have been divided by the weight for the number of rooms in the house, so their magnitudes are comparable to the ρs reported above. The coefficient on the principal-component asset index is 0.050. We rescale this coefficient by multiplying it by $\rho'\delta$ to make it comparable to the estimates where the assets are entered separately. In this formula ρ is the ratio of the bivariate correlations and δ is the vector of weights, with $\delta_j = score_j/\sigma_j$, where $score_j$ is the scoring factor for variable j, and σ_j is the standard deviation of that variable. The adjusted coefficient on the asset index is 0.098, which is over 40% smaller than the effect estimated when all the proxies are entered separately and their coefficients recombined. Indeed, the estimate of the first principal component from all 18 asset variables has more attenuation bias than each of our estimates that use only six of the asset variables.

Finally, we check to see whether our results are affected by the presence of covariates in the model. We estimate the covariate-adjusted estimator of ρ discussed in section IV,

[5] Although there are 21 separate variables, we label the two indicating toilet types and the two sources of drinking and nondrinking water as each being one, rather than two, proxies.

[6] The bootstrap procedure used 200 replications with 109,973 observations in each. The seed for the random number generator was set to 3,334,295.

TABLE 1.—MEASURING THE EFFECT OF WEALTH ON CHILDREN'S SCHOOL ATTENDANCE IN INDIA

		Proxy Set						Principal Components' Relative Weights
	ρ	1	2	3	4	5	6	
No. of rooms in house	1.000	0.009 (0.001)	0.011 (0.001)		0.013 (0.001)			1.000
Refrigerator	0.121	−0.035 (0.006)	−0.035 (0.006)		−0.029 (0.005)			0.546
Clock or watch	0.452	0.089 (0.003)	0.123 (0.003)		0.132 (0.003)			0.429
Type of toilet:								
Flush	0.275	0.046 (0.004)	0.074 (0.004)			0.113 (0.004)		0.077
Latrine	0.091	0.052 (0.004)	0.069 (0.004)			0.094 (0.004)		1.007
Sewing machine	0.291	0.039 (0.003)	0.067 (0.003)			0.098 (0.003)		0.922
VCR	0.038	−0.014 (0.008)	−0.033 (0.008)			−0.025 (0.008)		0.662
Radio	0.353	0.034 (0.003)	0.054 (0.003)				0.084 (0.003)	1.249
Drinking water from:								
Pump	−0.226	0.002 (0.011)	−0.032 (0.004)				−0.028 (0.004)	−0.154
Open source	−0.001	0.023 (0.012)	0.010 (0.005)				0.025 (0.005)	−0.694
Cooking fuel is wood, dung, or coal	−0.243	0.004 (0.004)	−0.008 (0.004)				−0.017 (0.004)	−1.097
Television	0.332	0.005 (0.004)		0.034 (0.004)	0.038 (0.004)			0.962
Nondrinking water from:								
Pump	−0.203	−0.014 (0.011)		−0.023 (0.004)	−0.038 (0.004)			−0.283
Open source	−0.028	0.021 (0.012)		0.029 (0.005)	0.007 (0.005)			1.099
Fan	0.414	0.022 (0.004)		0.053 (0.004)	0.097 (0.003)			1.216
Bicycle	0.188	0.019 (0.003)		0.036 (0.003)		0.046 (0.003)		0.934
Car	0.019	−0.062 (0.010)		−0.064 (0.010)		−0.062 (0.010)		−0.839
Kitchen in separate room	0.311	0.064 (0.003)		0.085 (0.003)		0.097 (0.003)		0.666
Motorcycle	0.129	−0.004 (0.005)		0.000 (0.005)			0.013 (0.005)	1.055
Electric lighting	0.447	0.114 (0.003)		0.137 (0.003)			0.162 (0.003)	0.616
Livestock	−0.179	−0.007 (0.003)		−0.009 (0.003)			−0.006 (0.003)	−0.573
Number of proxies		18	9	9	6	6	6	18
Estimated "wealth" effect		0.170 (0.003)	0.136 (0.003)	0.132 (0.003)	0.129 (0.003)	0.105 (0.002)	0.116 (0.003)	Unadjusted 0.050 (0.001) Rescaled 0.098 (0.001)

Note: Data are from the Demographic and Health Survey of India. Sample size is 109,973. The dependent variable is an indicator that the child is enrolled in school. The model also controls for the child's sex and age, the head of household's sex, age, and education, and the log family size. Standard errors are in parentheses. Construction of estimated wealth effects is described in the text.

using the residuals from regressions of the proxies and school enrollment indicator on all of the covariates. This procedure delivers an estimate of 0.194 (with a standard error of 0.006), which is even larger than our original estimate of 0.170. The covariate-adjusted estimate of ρ in our permanent-income example also leads to a larger estimate than that reported in figure 1, but in this case the adjusted estimate is quite imprecise and not statistically different from the unadjusted estimate.

VI. Conclusion

We have proposed a new estimator for the case where a researcher has multiple proxies for a single, unobserved independent variable. Numerous previous studies have dealt with the problem either by using the proxies one at a time, or by averaging or otherwise aggregating the proxies together and using that single measure as an independent variable. We show that attenuation bias is maximally reduced when the proxies are entered simultaneously in a multiple regression and the coefficients on them optimally combined after the fact to yield an estimate of the effect of the unobserved variable. To optimally weight the proxies prior to the regression requires knowing the variances and covariances between the error components in the proxies, information that is simply unavailable to the researcher. The improved performance of the estimator is due to the fact that the regression coefficients on the proxies precisely reflect this unknown information. This method is also more transparent than ad hoc index construction, because a reader who believes some proxies have independent effects on the dependent variable has the information available to create alternative estimates based on a subset of the proxies.

We have put off discussion of the asymptotic or finite-sample distribution of the our estimator. The need to estimate ρ introduces additional noise into the estimates that is

not present in an ad hoc index variable. Unreported Monte Carlo simulations suggest that this source of variance may not be particularly large and is probably outweighed by the large reduction in bias in the estimates themselves (see Lubotsky & Wittenberg, 2003). More generally, the analytic distribution of the estimators is quite difficult to compute, and researchers are probably better off using bootstrap methods to calculate the standard error of their estimates.

REFERENCES

Aigner, Dennis J., Cheng Hsiao, Arie Kapteyn, and Tom Wansbeek, "Latent Variable Models in Econometrics" (pp. 1321–1393), in Zvi Griliches and Michael D. Intriligator (Eds.), *Handbook of Econometrics,* Vol. II (Elsevier, 1984).

Blau, David M., "The Effect of Income on Child Development," this *REVIEW,* 81:2 (1999), 261–276.

Bollen, Kenneth A., *Structural Equations with Latent Variables* (New York: Wiley, 1989).

Bollinger, Christopher R., "Measurement Error in Human Capital and the Black-White Wage Gap," this *REVIEW,* 85:3 (2003), 578–585.

Case, Anne, Darren Lubotsky, and Christina Paxson, "Economic Status and Health in Childhood: The Origins of the Gradient," *American Economic Review* 92:5 (2003), 1308–1334.

Davidson, Russell, and James G. MacKinnon, *Estimation and Inference in Econometrics* (New York: Oxford University Press, 1993).

Dhrymes, Phoebus J., *Econometrics: Statistical Foundations and Applications* (New York: Springer, 1974).

Filmer, Deon, and Lant H. Pritchett, "Estimating Wealth Effects without Expenditure Data—Or Tears: An Application to Educational Enrollment in States of India," *Demography* 38:1 (2001), 115–132.

Fuller, Wayne A., *Measurement Error Models* (New York: Wiley, 1987).

Glaeser, Edward L., David I. Laibson, José A. Scheinkman, and Christine L. Soutter, "Measuring Trust," *Quarterly Journal of Economics* 115 (2000), 811–846.

Goldberger, Arthur S., "Structural Equation Methods in the Social Sciences," *Econometrica* 40:6 (1972), 979–1001.

Griliches, Zvi, "Economic Data Issues" (pp. 1466–1514), in Zvi Griliches and Michael D. Intriligator (Eds.), *Handbook of Econometrics,* Vol. III (Elsevier, 1986).

Herrnstein, Richard J., and Charles Murray, *The Bell Curve: Intelligence and Class Structure in American Life* (New York: The Free Press, 1994).

Jöreskog, Karl G., and Arthur S. Goldberger, "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," *Journal of the American Statistical Association* 70:351 (1975), 631–639.

Leamer, Edward E., "Model Choice and Specification Analysis," (pp. 285–330), in Zvi Griliches and Michael D. Intriligator (Eds.), *Handbook of Econometrics,* Vol. I (North-Holland, 1983).

Lubotsky, Darren, and Martin Wittenberg, "Interpretation of Regressions with Multiple Proxies," University of Illinois at Urbana-Champaign working paper (2003).

Mauro, Paolo, "Corruption and Growth," *Quarterly Journal of Economics* 110:3 (1995), 681–712.

Mayer, Susan E., *What Money Can't Buy: Family Income and Children's Life Chances* (Cambridge, MA: Harvard University Press, 1997).

Mazumder, Bhashkar, "Earnings Mobility in the U.S.: A New Look at Intergenerational Inequality," Federal Reserve Bank of Chicago working paper no. 2001–18 (2001).

Rao, C. Radhakrishna, and Helge Toutenburg, *Linear Models: Least Squares and Alternatives, 2nd ed.* (New York: Springer, 1995).

Solon, Gary, "Intergenerational Income Mobility in the United States," *American Economic Review* 82:3 (1992), 393–408.

Zimmerman, David J., "Regression toward Mediocrity in Economic Stature," *American Economic Review* 82:3 (1992), 409–429.

APPENDIX

Proofs

Lemma A.1.

1. $|A + aa'| = |A|(1 + a'A^{-1}a)$ if A is nonsingular.
2. $(A + ab')^{-1} = A^{-1} - \dfrac{A^{-1}ab'A^{-1}}{1+b'A^{-1}a}$ if $1 + b'A^{-1}a \neq 0$.

PROOF: Part 1 is Theorem A.16(x) in Rao and Toutenburg (1995, p. 358). Part 2 is Theorem A.18(iv) of Rao and Toutenburg (1995, p. 358).

Proof of Theorem 1: We have

$$\operatorname{plim} \hat{\beta} = \frac{\beta}{\delta'\rho}\left(1 - \frac{\delta'\Sigma_{UU}\,\delta}{\sigma_x^2(\delta'\rho)^2 + \delta'\Sigma_{UU}\,\delta}\right),$$

that is,

$$\operatorname{plim} (\rho'\delta) = \beta\left(1 - \frac{\delta'\Sigma_{UU}\,\delta}{\sigma_x^2(\delta'\rho)^2 + \delta'\Sigma_{UU}\,\delta}\right).$$

By contrast,

$$\operatorname{plim} b = \beta\sigma_x^2(\Sigma_{XX})^{-1}\rho.$$

Because $\Sigma_{XX} = \Sigma_{UU} + \sigma_x^2\rho\rho'$, we can apply lemma A.1 It follows that

$$\textstyle\sum_{XX}^{-1}\rho = \dfrac{\sum_{UU}^{-1}\rho}{1 + \sigma_x^2\rho'\sum_{UU}^{-1}\rho}. \tag{A-1}$$

Hence

$$\operatorname{plim} \rho b = \beta\left(1 - \frac{1}{1 + \sigma_x^2\rho'\sum_{UU}^{-1}\rho}\right).$$

Observe that we can apply lemma A.1 again to show that

$$\frac{1}{1 + \sigma_x^2\rho'\sum_{UU}^{-1}\rho} = \frac{\left|\sum_{UU}\right|}{\left|\sum_{XX}\right|}. \tag{A-2}$$

We want to compare $1/\sigma_x^2\rho'\,\Sigma_{UU}^{-1}\rho + 1$ and $\delta'\Sigma_{UU}\,\delta/[\sigma_x^2(\delta'\rho)^2 + \delta'\,\Sigma_{UU}\,\delta]$, so we need to show that

$$\rho'\,\textstyle\sum_{UU}^{-1}\rho \geq \dfrac{(\delta'\rho)^2}{\delta'\sum_{UU}\delta}$$

for any nonzero choices of ρ and δ.

Because Σ_{UU} is a nonsingular covariance matrix, by the spectral theorem for symmetric matrices it can be decomposed as

$$\textstyle\sum_{UU} = PDP'$$

where P is an orthogonal matrix of eigenvectors

$$P = [p_1\, p_2 \cdots p_k]$$

and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$ is the matrix of eigenvalues, with $\lambda_i > 0\ \forall_i$. This is equivalent to writing

$$\textstyle\sum_{UU} = \lambda_1 p_1 p_1' + \cdots + \lambda_k p_k p_k',$$

and it follows that

$$\textstyle\sum_{UU}^{-1} = \dfrac{1}{\lambda_1}p_1 p_1' + \cdots + \dfrac{1}{\lambda_k}p_k p_k'.$$

Hence

$$\rho' \sum_{UU}^{-1} \rho = \sum_i \frac{1}{\lambda_i}(p_i' \rho)^2,$$

$$\delta' \sum_{UU} \delta = \sum_i \lambda_i (p_i' \delta)^2.$$

Now let $p_i'\rho = w_i$ and $p_i'\delta = v_i$. Correspondingly define the vectors w and v as

$$w = P'\rho, \quad v = P'\delta.$$

Note that

$$\delta'\rho = v'w$$

(because P is orthogonal), that is,

$$\begin{aligned}(\delta'\rho)^2 &= (\sum_i v_i w_i)^2 \\ &\le \left(\sum_i \sqrt{\lambda_i} \left| v_i \right| \left| \frac{1}{\sqrt{\lambda_i}} \right| \left| w_i \right| \right)^2 \\ &\le (\sum_i \lambda_i v_i^2)\left(\sum_i \frac{1}{\lambda_i} w_i^2 \right) \qquad \text{(Cauchy-Schwarz inequality)} \\ &= (\delta' \sum_{UU} \delta)(\rho' \sum_{UU}^{-1} \rho).\end{aligned}$$

Equality holds only if

$$\sqrt{\lambda_i} v_i = \frac{c}{\sqrt{\lambda_i}} w_i$$

for some real number c, that is,

$$\begin{aligned}\lambda_i v_i &= c w_i \\ \delta &= cPD^{-1}P'\rho \\ &= c \sum_{UU}^{-1} \rho.\end{aligned}$$

PROOF OF PROPOSITION 2: We have

$$\text{plim } b_i = e_i' \sum_{XX}^{-1} \beta \rho \sigma_x^2,$$

and by using equations (A-1) and (A-2),

$$e_i'(\sum_{XX})^{-1} \rho \sigma_x^2 = \frac{|\sum_{UU}|}{|\sum_{XX}|}(\sigma_x^2 e_i' \sum_{UU}^{-1} \rho).$$

Now

$$\sum_{UU}^{-1} = \frac{1}{|\sum_{UU}|} \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix},$$

where S_{ij} is the ij^{th} cofactor of Σ_{UU}. Consequently

$$e_i' \sum_{UU}^{-1} = \frac{1}{|\sum_{UU}|}[S_{i1}\ S_{i2} \ldots S_{ik}],$$

that is,

$$e_i' \sum_{UU}^{-1} \rho = \frac{1}{|\sum_{UU}|} \sum_j S_{ij} \rho_j.$$

The summation $\Sigma_j\ S_{ij}\rho_j$, however, is identical to the value of the determinant if row i of the matrix Σ_{UU} were replaced by ρ', that is,

$$\sum_j S_{ij}\rho_j = |\sum_{UU}^{\rho(i)}|.$$

Consequently

$$e_i'(\sum_{XX})^{-1} \rho \beta \sigma_x^2 = \frac{\beta \sigma_x^2 |\sum_{UU}^{\rho(i)}|}{|\sum_{XX}|}$$

PROOF OF PROPOSITION 3: Assume that the true model is given by

$$y = x^*\beta + Z\gamma + \varepsilon,$$

where Z is a matrix of covariates. According to the FWL theorem (Davidson & MacKinnon, 1993, p. 19ff.), the regression coefficients in this multiple regression are numerically identical to the coefficients in the regression

$$M_Z y = M_Z x^* \beta + M_Z \varepsilon,$$

where M_Z is the matrix $I - Z(Z'Z)^{-1}Z$, that is, the matrix that generates the residuals in the regression of any n-vector on Z. If we rewrite this latter regression as

$$e_y = e_x^* \beta + \upsilon, \tag{A-3}$$

where e_y is the vector of residuals $M_Z y$, e_x^* is the vector of the unobservable residuals $M_z x^*$, and v is a vector of error terms, we can apply the results of the previous sections directly to this regression. In particular, the matrix e_X of observable residual proxies $M_Z X$ will serve as proxies for the unobservable e^*. Note that

$$\begin{aligned} M_Z x_j &= \rho_j M_Z x^* + M_Z u_j, \\ e_j &= \rho_j e^* + w_j, \end{aligned} \tag{A-4}$$

so that the our estimator applied to the regression (A-4) is just $\rho' b$, where b is the OLS coefficient in the regression of e_y on $M_Z X$. This, however, is numerically equal to the OLS coefficient in the regression of y on the matrix of proxies X and the matrix of covariates Z, that is,

$$y = Xb + Z\gamma + \varepsilon.$$

Furthermore, $\rho_j = \text{cov}(y,x_j)/\text{cov}(y,x_1)$ is still a consistent estimator of ρ_j provided that Z is uncorrelated with u_j and u_1. In other words, our estimator $\hat{\rho}'b$ will again provide the attenuation bias minimizing combination of the proxies on condition that Z is uncorrelated with the measurement error in the proxies.

One difference, however, is that the attenuation bias in this case is not given by the formula

$$\text{plim } \hat{\rho}'b = \beta\left(1 - \frac{|\sum_{UU}|}{|\sum_{XX}|}\right),$$

but by

$$\text{plim } \hat{\rho}'b = \beta\left(1 - \frac{|\sum_{UU}|}{|\sum_{M_Z X M_Z X}|}\right),$$

where $\Sigma_{M_Z X M_Z X}$ is the covariance matrix of $M_Z X$. This result is obvious on applying Theorem 1 to equation (A-3), bearing in mind the formula for the proxies given in equation (A-4) and noting that $M_Z u_j = u_j$ by the assumption that Z is uncorrelated with the errors. In this case therefore the numerator of the original attenuation term $|\Sigma_{UU}|/|\Sigma_{XX}|$ is not affected, while the denominator is reduced. That is, attenuation bias increases.

If Z is correlated with any of the measurement error terms, then a consistent estimator of ρ can still be obtained, using the approach of the previous section applied to equations (A-3) and (A-4). In this case $\rho_j = \text{cov}\ (e_y, e_j)/\text{cov}\ (e_y, e_1)$ is a consistent estimate; that is, the correlation between y and the proxies that works through the measurement error term would first have to be removed. Note that the original formula still applies in the special case where the covariate is a constant.

The attenuation bias formula would now be given by

$$\text{plim}\ \hat{\rho}'b = \beta\left(1 - \frac{\left|\sum_{M_ZUM_ZU}\right|}{\left|\sum_{M_ZXM_ZX}\right|}\right),$$

where $\Sigma_{M_ZUM_ZU}$ is the covariance matrix of M_ZU. In this case attenuation may, in fact, be reduced at the cost of increasing the bias on the coefficient of the correctly measured variable. Furthermore, the estimator of ρ is likely to be noisier.

To show the effect of the proxies on the correctly measured covariates, we can apply the FWL theorem in reverse. The coefficient γ in the multiple regression is identical to the coefficient of γ in the regression of the residuals $M_x{*}y$ on the residuals $M_x{*}Z$. In our case, however, due to attenuation bias, the residuals M_Xy would not have been entirely purged of the effect of x^*, due to the measurement error. Similarly, the residuals M_XZ would still contain some of the effect of x^*. Consequently the estimate of γ would represent not only the direct effect of Z on y but also some of the residual effect of x^* on y that has not adequately been controlled for by the proxies.

Another way of viewing this case is to simply note that the regression results in the full proxy variable regression

$$y = Xb + Z\gamma + \varepsilon$$

are numerically identical to the results obtained if we ran the regression on our index, that is,

$$y = x^pb^p + Z\gamma + \varepsilon.$$

Because this regression fits into the mold of the classical measurement error literature, the results for γ follow.

If Z is correlated with any of the measurement error terms, then the residuals M_XZ would remove some of the direct effects of Z, so that the bias on γ would be exacerbated. ■