Interpretation of Regressions with Multiple Proxies[*]

Darren Lubotsky
University of Illinois at Urbana–Champaign
lubotsky@uiuc.edu

Martin Wittenberg
University of the Witwatersrand
wittenbergm@sebs.wits.ac.za

June 2002

Abstract

We consider the situation in which there are multiple proxies for one unobserved explanatory variable in a linear regression and provide a procedure by which the coefficient of interest can be extracted "*post hoc*" from a multiple regression in which all the proxies are used simultaneously. This *post hoc* estimator is strictly superior in large samples to coefficients derived using any index or linear combination of the proxies that is created prior to the regression. To use an index created from the proxies that extracts the largest possible signal from them requires knowledge of information that is not available to the researcher. Using the proxies simultaneously in a multiple regression delivers this information and the researcher then simply combines the coefficients in a known way to obtain the estimate of the effect of the unobserved factor. This procedure is also much more robust than *ad hoc* index construction to departures from the assumption of an underlying common factor. We provide some Monte Carlo simulations and applications to existing empirical problems to show that the reduction in attenuation bias can be non-negligible, even in finite samples.

JEL Classification: C1, C31
Keywords: Proxy variables, measurement error, index construction.

[*] We thank Anne Case, Angus Deaton, Bo Honoré, Hashem Pesaran, Elie Tamer and seminar participants at the African Econometrics Conference, the North American Summer Meeting of the Econometric Society, and Princeton University for their comments and suggestions. We are naturally responsible for any errors. Much of this work was completed while both authors were affiliated with the Research Program in Development Studies at Princeton University. Martin Wittenberg acknowledges financial support from Investec Bank through their grant to Econometric Research Southern Africa and we both acknowledge financial support from the John D. and Catherine T. MacArthur Foundation through their Network on Poverty and Inequality in a Broader Perspective.

1 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. While a great deal of attention has been paid to 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 there is more information available, the problem of attenuation should be reduced. The fundamental question is really 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:

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. Permanent income is not observed; instead, observed income in any year includes transitory components representing luck, measurement error, and other unanticipated shocks to income. If parents' investment in their children is a function of permanent income, then a regression of children's outcomes on parents' observed income will understate the true effect of permanent income on outcomes. 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 this approach is it assumes that income earned at different points in the lifecycle 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 might do even better than a simple average, with more weight given to income earned in later years. We will show below how such weights can be computed and that a more optimal use of annual earnings data increases the estimated effect of parents' permanent income on children's reading test scores by over 30 percent.

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 or educational attainment, therefore, 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 their 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 plus a function of tastes or other characteristics of the household, the first 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.

The two examples 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 sources of noise, however. Glaeser, Laibson, Scheinkman and Soutter (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 indices 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. Since he believes many of these indices measure the same underlying phenomena, he averages the indices 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.

There are several considerations underlying 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 wealth on education from the coefficients on variables indicating ownership of a television or the availability of running water in the home.

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 sceptical 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. The main theoretical results are in section 3. We provide some simulation evidence in section 4 and then return to the examples described above in section 5. We conclude by pointing to a number of open questions, and an appendix contains the proofs of our main results.

2 The basic problem

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

$$y_i = \beta x_i + \varepsilon_i \tag{1a}$$

$$x_{1i} = x_i + u_{1i} \tag{1b}$$

$$x_{2i} = \rho_2 x_i + u_{2i} \tag{1c}$$

where β, relating y_i and x_i in equation 1a, is the parameter of interest. We assume that x_i is unobserved, but that we have the two proxies x_{1i} and x_{2i}. Furthermore we will make the assumption that u_{1i} and u_{2i} are independent of x_i and ε_i; that is, the proxy variables do not have an independent effect on y_i.

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 $\sigma_x^2 = var(x)$ and $\sigma_1^2 = var(u_1)$. The parameter β is not identified. As Aigner, Hsiao, Kapteyn and Wansbeek (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 $cov(u_1, u_2) = 0$, 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 sub–set 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 only with even more stringent assumptions. Not only do we need to impose orthogonality between the error variances u_1 and u_2, but we also need to 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 combined 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 there is at least one more relationship 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 series of observable variables, i.e. equations 1b and 1c are replaced by

$$x_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_3$$

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 to best 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

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

and the covariance matrix of u_1 and u_2 by

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

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

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

Since 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

$$\overline{x}_i = x_i + \overline{u}_i$$

with $var\,(\overline{u}_i) = \frac{1}{4}\left(\sigma_1^2 + \sigma_2^2 + 2\sigma_{12}\right)$. 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 $\delta_1 = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}$. The variance in this case will be $\frac{\sigma_2^2\sigma_1^2 - \sigma_{12}^2}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}$, so that the estimate of β with the minimum attenuation bias is given asymptotically by

$$b^* = \beta\frac{\sigma_x^2}{\sigma_x^2 + \frac{\sigma_2^2\sigma_1^2 - \sigma_{12}^2}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}} \tag{3}$$

Unfortunately, we do not know the variances and covariance of u_1 and u_2, and thus 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

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

It is straightforward to check that

$$b_1 = \beta\frac{\sigma_x^2\left(\sigma_2^2 - \sigma_{12}\right)}{\sigma_x^2\sigma_1^2 + \sigma_x^2\sigma_2^2 - 2\sigma_x^2\sigma_{12} + \sigma_1^2\sigma_2^2 - \sigma_{12}^2} \tag{4a}$$

$$b_2 = \beta\frac{\sigma_x^2\left(\sigma_1^2 - \sigma_{12}\right)}{\sigma_x^2\sigma_1^2 + \sigma_x^2\sigma_2^2 - 2\sigma_x^2\sigma_{12} + \sigma_1^2\sigma_2^2 - \sigma_{12}^2} \tag{4b}$$

This does not look very promising, but note that

$$\begin{aligned} b_1 + b_2 &= \beta \frac{\sigma_x^2 \left(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\right)}{\sigma_x^2\sigma_1^2 + \sigma_x^2\sigma_2^2 - 2\sigma_x^2\sigma_{12} + \sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ &= b^* \end{aligned} \tag{5}$$

so that adding up the coefficients of the two variables yields an estimate that is precisely equal to the optimal weighting 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 – that the attenuation bias is always smallest when all the proxies are used in a multiple regression. We need to proceed with some care, however, in the situation where $\rho_2 \neq 1$. In that case we note that a simple average of the variables is

$$\overline{x}_i = \overline{\rho} x_i + \overline{u}_i$$

with $\overline{\rho} = \frac{1+\rho_2}{2}$. To avoid this unknown rescaling of the latent variable x, we will generally want to take a weighted sum of the individual coefficients from the multiple regression.

3 The general case

We now assume that we have k proxies, with

$$x_j = \rho_j x + u_j \tag{6}$$

We assume that $cov\left(u_j, \varepsilon_i\right) = 0$, $cov\left(u_j, x\right) = 0$ for all j, but that the covariance matrix of u_js is unrestricted, i.e.

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

where $U = \begin{bmatrix} u_1 & u_2 & \ldots & u_k \end{bmatrix}$.

3.1 Identification

We note that as it stands the ρ_j terms are not identified. Rescaling the ρs and β would result in the same observations. 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.[1]

The available information is contained in the covariance matrix

$$\Sigma_{ZZ} = E\left(Z'Z\right) = \begin{bmatrix} \beta^2\sigma_x^2 + \sigma_\varepsilon^2 & \beta\sigma_x^2 & \beta\rho_2\sigma_x^2 & \cdots & \beta\rho_k\sigma_x^2 \\ \beta\sigma_x^2 & \sigma_x^2 + \sigma_1^2 & \rho_2\sigma_x^2 + \sigma_{12} & \cdots & \rho_k\sigma_x^2 + \sigma_{1k} \\ \beta\rho_2\sigma_x^2 & \rho_2\sigma_x^2 + \sigma_{12} & \rho_2^2\sigma_x^2 + \sigma_2^2 & \cdots & \rho_2\rho_k\sigma_x^2 + \sigma_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta\rho_k\sigma_x^2 & \rho_k\sigma_x^2 + \sigma_{1k} & \rho_2\rho_k\sigma_x^2 + \sigma_{2k} & \cdots & \rho_k^2\sigma_x^2 + \sigma_k^2 \end{bmatrix} \tag{7}$$

where $Z = \left[\begin{array}{ccccc} y & x_1 & x_2 & \ldots & x_k \end{array}\right]$.

There are altogether $\frac{k(k+1)}{2}$ unknown parameters in Σ_{UU}, $k-1$ parameters in ρ and the parameters β, σ_x^2 and σ_ε^2. Since there are altogether $\frac{(k+1)(k+2)}{2} + 1$ unknown parameters in Σ_{ZZ}, but only $\frac{(k+1)(k+2)}{2}$ pieces of observable information, we are therefore 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:

$$\rho_j = \frac{cov\left(y, x_j\right)}{cov\left(y, x_1\right)} \tag{8}$$

As noted above there are several ways in which we could achieve identification of β: any restriction on the covariance matrix of $X = \left[\begin{array}{cccc} x_1 & x_2 & \ldots & x_k \end{array}\right]$ 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. Since we have an estimator for ρ_j this is easily achieved. We have

$$\beta = \frac{cov\left(y, x_h\right)}{cov\left(x_j, x_h\right)} \frac{cov\left(y, x_j\right)}{cov\left(y, x_1\right)}$$

The first term is the "instrumental variables" estimator while the second is the GMM estimator of ρ_j.

[1] Other normalizations are possible; e.g. we might fix $||p|| = 1$. A problem with such a normalization is that the resulting scale of the coefficient is difficult to interpret.

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.

3.2 Minimizing attenuation bias

Let

$$x^\delta = X\delta \tag{9}$$

be any linear combination of the proxy variables where $X = \begin{bmatrix} x_1 & x_2 & \dots & x_k \end{bmatrix}$. By assumption

$$X = x\rho' + U$$

where $\rho' = \begin{bmatrix} 1 & \rho_2 & \dots & \rho_k \end{bmatrix}$. It follows that

$$x^\delta = x\rho'\delta + U\delta$$

Unless $\rho'\delta = 1$ this will involve a rescaling, so we will want to multiply our final estimates by $\rho'\delta$ to make all results comparable. The generalization of the discovery we made in the two proxy case with $\rho_2 = 1$ is contained in the following theorem:

Theorem 1 *Let $b^\delta = \widehat{\beta}\,(\rho'\delta)$ be the rescaled OLS estimate in the regression of y on x^δ, b be the OLS estimate in the regression of y on X, and $\widehat{\rho}$ be the GMM estimate of ρ. Then*

$$\begin{aligned} \mathit{plim}\ b^\delta &= \beta\left(1 - \frac{\delta'\Sigma_{UU}\delta}{\sigma_x^2\left(\delta'\rho\right)^2 + \delta'\Sigma_{UU}\delta}\right) \\ \mathit{plim}\ \widehat{\rho}'b &= \beta\left(1 - \frac{|\Sigma_{UU}|}{|\Sigma_{XX}|}\right) \\ &= \beta\left(1 - \frac{1}{\sigma_x^2\rho'\Sigma_{UU}^{-1}\rho + 1}\right) \end{aligned} \tag{10}$$

where Σ_{XX} is the covariance matrix of X. Furthermore for every $\delta \neq 0$ we have

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

Equality holds only if

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

for some $c \neq 0$.

We note that the appropriate way of aggregating up the coefficients in the multiple regression is given by the *"post hoc"* estimator

$$b^p = \widehat{\rho}' b = \sum_{j=1}^{k} \frac{cov\,(y, x_j)}{cov\,(y, x_1)} b_j \qquad (11)$$

where b_j is the coefficient on x_j in the multiple regression. We use the term *post hoc* both because the estimation happens after the event (the regression), but also because it can be seen as a rationalization of the data: from the k different regression coefficients on the proxies, the *post hoc* estimator gives a way to interpret how changes in the underlying unobserved variable x effect the dependent variables. 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.

The theorem proves that no linear combination of the proxies will achieve a greater reduction in attenuation bias than our process of *post hoc* inference. 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 essence the multiple regression provides the appropriate reweighting of the variables to minimize the error variance of the aggregate set of proxies.

We note that the formula in equation 10 provides the natural generalization of 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).

3.3 *Post hoc* inference and index construction

Our procedure obviously depends on the validity of the underlying assumptions. If 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.

One attractive feature is that it is possible to provide the reader with the estimates of the ρs and the reader can then assess how plausible the assumption of the "common factor" is. 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 on to it. Furthermore, given the ρs, it is possible to provide different estimates of β, depending on whether particular proxies are viewed as having independent effects or not.

Strictly speaking our procedure should therefore not be viewed as an *estimation* but as an *interpretation* procedure. Equation 11 tells us how to interpret the coefficients of a multiple regression under the null hypothesis that there is a common underlying factor which is generating the separate coefficients.

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 (11). 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 *post hoc* as well:

$$x^p = \frac{1}{b^p} \sum_{j=1}^{k} x_j b_j \tag{12}$$

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^p as the coefficient in the regression.[2]

This *post hoc* 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, by comparing the correlation structure between other variables and an index representing wealth 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 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^p 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, i.e.*

$$b_i = e_i' \left(X'X\right)^{-1} X'y$$

where e_i is the unit vector with one in the i-th position. Then

$$plim\ b_i = \beta \frac{\sigma_x^2 \left|\Sigma_{UU}^{\rho(i)}\right|}{\left|\Sigma_{XX}\right|} \tag{13}$$

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

[2] We have $\delta = \frac{1}{b^p} b$ with $b = \left(X'X\right)^{-1} X'y$. Consequently $\delta'\rho = 1$ (since $b^p = b'\rho$) and $\left(\delta'X'X\delta\right)^{-1} \delta'X'y = b^p$.

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

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

Several additional points follow from this result. Firstly, if $\beta = 0$, then every single proxy coefficient must be zero. This means that the hypothesis that $\beta = 0$ is testable as a joint hypothesis on all the proxies. Indeed, it is also testable on the sum of the proxies.

Secondly, it follows from our proof about the bias in $\rho' b$ that

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

There are therefore $k+1$ terms in the denominator of equation 13. If the proxies are all of similar quality, i.e. if the ρs and error variances are not vastly dissimilar, then the individual coefficients should be of the order of $\frac{\beta}{k}$, i.e. as more proxies are added, the individual coefficients should tend to zero. 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.

3.4 The impact of other 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. Both of these results also apply to the bias in the *post hoc* estimator when multiple proxies for an unobserved variable are included in the regression. The coefficients on the covariates will be biased as well, the magnitude of which depends on the covariances between the covariates, the unobserved variable, and the measurement error components in the proxies (u_j above).

In this case, however, it is particularly important to be concerned about the correlation between the measurement error component and the covariates. Adding in proxies that absorb the effects of the covariates instead of proxying for the latent variable would be particularly damaging. An important trade–off exists, therefore, in adding additional 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. The *post hoc* interpretation procedure should therefore not be taken as license to throw any and all variables into the regression. Ideally, the proxies should be correlated with x and their measurement error components should be orthogonal to the other explanatory variables in the regression.

3.5 Biased instrumental variables estimation

The results above indicate that our *post hoc* approach 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 the *post hoc* approach. Indeed we can investigate under which circumstances this is likely to be the case.

In the two variable case given in equations 1a-1c, if we use x_2 as an instrument for x_1, then asymptotically

$$\begin{aligned} b_{iv} &= \frac{cov\left(x_2, y\right)}{cov\left(x_2, x_1\right)} \\ &= \beta \frac{\sigma_x^2}{\sigma_x^2 + \frac{\sigma_{12}}{\rho_2}} \end{aligned} \tag{14}$$

It is obvious that the smaller the covariance between u_1 and u_2 is, the smaller the asymptotic bias. Unlike in the least squares case, however, the direction of the bias depends on the sign of σ_{12}. There is no longer a guarantee that the estimate is a lower bound on the true value.

If we let $\rho_2 = 1$ we can compare the absolute value of the IV bias to the bias in the estimate of b^* (equation 3). Instrumental variables will yield a smaller absolute bias if, and only if

$$|\sigma_{12}| \leq \frac{\sigma_2^2\sigma_1^2 - \sigma_{12}^2}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}$$

If $\sigma_{12} > 0$ we get the condition

$$0 \leq \sigma_2^2\sigma_1^2 - \left(\sigma_1^2 + \sigma_2^2\right)\sigma_{12} + \sigma_{12}^2$$

This quadratic in σ_{12} is guaranteed to have real roots. The condition will be satisfied if, and only if

$$0 \leq \sigma_{12} \leq \min\left(\sigma_1^2, \sigma_2^2\right) \text{ or } \sigma_{12} \geq \max\left(\sigma_1^2, \sigma_2^2\right)$$

The latter condition is irrelevant, since the positive definiteness of Σ_{UU} implies that $\sigma_{12} \leq \frac{\sigma_1^2+\sigma_2^2}{2}$. The condition therefore simplifies to

$$\sigma_{12} \leq \min\left(\sigma_1^2, \sigma_2^2\right)$$

It follows that if the error variances are positively correlated, biased IV will be superior provided the error variances are of similar magnitudes

If $\sigma_{12} < 0$ we get the condition

$$3\sigma_{12}^2 - \left(\sigma_1^2 + \sigma_2^2\right)\sigma_{12} - \sigma_2^2\sigma_1^2 \leq 0$$

which gives

$$\frac{\sigma_1^2 + \sigma_2^2}{6} - \frac{\sqrt{\sigma_1^4 + 14\sigma_1^2\sigma_2^2 + \sigma_2^4}}{6} \leq \sigma_{12} \leq 0$$

In the particular case where the error variances are equal, this condition is equivalent to $\sigma_{12} \geq -\frac{1}{3}\left(\sigma_u^2\right)$ where σ_u^2 is the common error variance.

Combining the two cases we find that biased IV is superior if, and only if,

$$\frac{\sigma_1^2 + \sigma_2^2}{6} - \frac{\sqrt{\sigma_1^4 + 14\sigma_1^2\sigma_2^2 + \sigma_2^4}}{6} \leq \sigma_{12} \leq \min\left(\sigma_1^2, \sigma_2^2\right) \tag{15}$$

It is evident that in many situations biased IV will improve on ordinary least squares. However, it is impossible to determine which is the better estimator without knowing the magnitudes of error variances and covariance.

If we have more than two proxies, there is yet further scope for improvement. In this case, however, the problem is to find the linear combination of proxies where the error is least correlated with the error in x_1. The conventional two-stage least squares estimate (using all the proxies as instruments for x_1) will definitely fare badly in this regard, since they will seek to explain not only the part of x_1 which is correlated with x, but also the error term u_1.

We can put the problem more formally as follows: let $\gamma = [0, \gamma_2, \ldots, \gamma_k]'$ be a column vector of real numbers with $\gamma_1 = 0$, so that $X\gamma$ is an arbitrary linear combination of $x_2 \ldots x_k$. Using $X\gamma$ as an instrument for x_1 yields asymptotically

$$\begin{aligned} b_{iv}^{\gamma} &= \beta\frac{\sigma_x^2\rho'\gamma}{\sigma_x^2\rho'\gamma + \sum_{i=2}^k \gamma_i\sigma_{1i}} \\ &= \beta\frac{\sigma_x^2}{\sigma_x^2 + \frac{1}{\rho'\gamma}\sum_{i=2}^k \gamma_i\sigma_{1i}} \end{aligned}$$

The "best" instrument is that linear combination which yields the smallest absolute value of $\frac{1}{\rho'\gamma}\sum_{i=2}^{k}\gamma_i\sigma_{1i}$. We can rewrite this expression as $\sum_{i=2}^{k}\omega_i\frac{\sigma_{1i}}{\rho_i}$, where $\omega_i = \frac{\rho_i\gamma_i}{\sum\rho_i\gamma_i}$, so that we are looking for a "weighted average" of the terms $\frac{\sigma_{1i}}{\rho_i}$. Unfortunately neither the weights nor the terms need to be positive, so the direction of the bias is indeterminate. Furthermore, the information necessary to compute the instrument is not estimable. Even if such an instrument could be constructed, it would yield a smaller absolute bias only if

$$\left|\frac{1}{\rho'\gamma}\sum_{i=2}^{k}\gamma_i\sigma_{1i}\right| \leq \frac{1}{\rho'\Sigma_{UU}^{-1}\rho} \tag{16}$$

With the information available to a researcher it is impossible to construct the appropriate instrument, ascertain whether the condition above is satisfied, or what the direction of the bias is. Consequently reporting the *post hoc* least squares estimates jointly with any IV estimates is likely to be a preferred strategy.

4 A Monte Carlo investigation of finite sample properties

The results above are all asymptotic. In finite samples we are faced by a number of problems. Firstly, we need to estimate the ρs, and this will increase the noise in our procedure. Secondly, there are trade-offs between degrees of freedom lost from including too many proxies and the increased precision gained by putting them all in.

To investigate these issues we run a Monte Carlo simulation. While the relative performance of different estimators will depend on the parameters we set for the simulations, by knowing the true data generating process we are able to assess the overall bias in all estimates. More importantly, we can compare the performance of the *post hoc* estimator to some of the other approaches commonly used. In our simulations we find that the coefficient from the *post hoc* procedure is about 20 percent larger than the coefficients from other procedure, which eliminates about three–quarters of the overall attenuation bias. In section 5 we take a different approach and compare the estimators using actual datasets and again find significant increases in the coefficient estimates.

For each of 100 runs in the Monte Carlo simulation, we draw 100 independent observations on x and ε from $\mathrm{N}(0,2)$, and then create y equal to $10 + 100 * x + \varepsilon$. We generate 20 proxy variables for x determined by $x_j = \rho_j x + u_j$ with the properties that each proxy has a different correlation (ρ_j) with the unobserved factor x, has a different error variance (u_j), and the error components are correlated with each other. Specifically, $\rho_1 = 1$ and for $j = 2$ to 20, ρ_j is randomly drawn from

U$(0,2)$. Var$(u_j) = 1.1^{j-1}$ and E$(u_j u_k) = 0.5^{k-j}\sigma_j\sigma_k$ for $j \neq k$. ρ and the covariance structure of the u_js are fixed for the 100 runs. We randomly reorder the proxies before each run to avoid systematically changing the quality of the proxies as we successively add them to the model. To ensure that our inferences about different estimators are not driven by the particular ρ vector we drew, the 100 simulations are repeated ten times with ten different draws of ρ, and the results averaged together.

Table 1 displays the mean and standard deviation of five estimators. Each row indicates the number of proxies used in the model. The mean square errors of the final model, which uses all 20 proxies, is given at the bottom of the table. The first column shows results when y is regressed on all of the proxies and the coefficients are averaged together weighted by the true value of ρ. In the second column the coefficients are weighted by an estimate of ρ, $\hat{\rho}_j = \frac{cov(y,x_j)}{cov(y,x_1)}$. Since $\rho_1 = 1$ the estimates with only one proxy are identical, 66.77, which is biased downwards by 33 percent. As more proxies are added, the estimates based on the true ρ and the estimated ρ remain very close to each other; with 10 proxies the bias is about 10 percent in each and with 20 proxies it 5.7 and 4 percent. Since ρ can be consistently estimated, the only cost in having to do so is the additional imprecision in the estimate. With 20 proxies the standard error rises from 2.8 to 7.2 as a result of having to estimate ρ. Note that the bias is slightly smaller in the second column than in the first.

The third through fifth columns implement alternative estimators that have appeared in the literature. In these we regress y on the unweighted average of the proxies, on the average of proxies after they have been standardized to have a mean of zero and unit variance, and on the first principal component of the proxies. We rescale the results for these estimators as appropriate to be comparable to those in the first two columns. All three of these estimator perform considerably poorer than the *post hoc* estimators in the first two columns. Although additional proxies improve each estimator, with 20 variables the bias in each is 18.5, 20.6, and 14.5 percent respectively. Since these estimates do not require the additional estimation of the weights placed on each proxy, there is a small gain in the precision of the estimates, but as the last row shows, the mean squared error is still between 3.5 and 7 times as large as that in second column. We conclude from these simulations that in situations where one cannot be confident that the proxies are of similar quality (i.e. similar ρs and error variances), the reduction in bias from aggregating the coefficients on the proxies can be considerable, compared to entering a single, essentially arbitrary combination of the proxies in the regression.

5 Applications to existing research

We illustrate the procedure with two empirical examples. Though we do not know the true data generating process and thus 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 effect of a family's permanent income on children's performance on a reading comprehension test. Permanent income is not observed and we instead have panel data on annual family annual 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.

A standard model of (log) permanent income specifies observed income at age t as being a function of unobserved permanent income (y^p), lifecycle 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. u_t reflects deviations from age–adjusted permanent earnings, which may be serially correlated and heteroscedastic.

If parents' investment in their children is a function of their permanent income, then the correlation between observed income and child outcomes understates the true correlation. To circumvent this problem, a general practice in the literature has been to average annual income over several years. See, for example, Blau (1999), Case, Lubotsky and Paxson (forthcoming), Mayer (1997), Solon (1992), and Zimmerman (1992).

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 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 poses a difficulty for including annual incomes separately in the regression. Therefore, we work with two–year averages of family income taken when the mothers where between the ages of 22 and 39. Our sample contains 7898 children–year observations of those aged six to fourteen 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 when income between ages 22 and 31 is used, and finally to 1.6 when income through age 39 is used.

Next we include family income from different periods in the regression separately and average the coefficients, first unweighted and then weighted by the GMM estimate of ρ. The estimates of ρ are also given in the table and they show a steady rise over the lifecycle, consistent with earlier incomes understating permanent income. The unweighted average of the regression coefficients assumes that the correlation between the test score and family income in all periods are equal, a restriction the data in fact soundly reject. Compared to entering average income in the regression, the unweighted average of the income coefficients still allows the variance of the transitory component of income to vary over the lifecycle. We find that the unweighted average of the income coefficients produces a total effect that is in some cases 27 percent larger than the effect of average income.

Finally, we optimally weight the income coefficients and this leads to a substantial rise in the relationship between children's test scores and permanent income.[3] Using family income when the mother is aged 22 to 39, the effect from using the optimally weighted coefficients is 2.1, compared to only 1.6 when income is averaged prior to the regression, an increase of 30 percent. Our optimal estimator implies that a 50 percent rise in permanent income would lead to a 1.1 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.

Our second empirical application reexamines Filmer and Pritchett's attempt to estimate the effect of household wealth on Indian children's propensity to be enrolled in school. The catch is that the Demographic and Health Survey of India does not contain any income or wealth data,

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

but it does 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, indicators for whether the household has a refrigerator, clock or watch, sewing machine, VCR, radio, television, fan, bicycle, car, motorcycle, electric lighting, a flush toilet or latrine, 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 2 displays our estimates of ρ, the ratios of the bivariate correlations between each asset and the school enrollment indicator, to the correlation between the number of rooms in the house and the indicator. Thus the units of our unobserved wealth index is the number of rooms in the house. Assets that are more common among poorer households, such as obtaining water from pumps or an open source, using wood, cow dung, or coal as cooking fuel, or owning livestock, have negative correlations with children's school enrollment. The number of rooms in the house has the largest correlation. The assets one would associate with the relatively best–off in the data – having a car, VCR, or refrigerator – are only owned by a small proportion of the household, and thus do very poorly in accounting for enrollment rates among the whole population.

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_js) than their raw correlation with school enrollment (the $\rho_j s$). 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 4, 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.[4] To see how

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

sensitive the estimate is to using fewer 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 less 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.

The last column of Table 2 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 estimated below. The coefficient on the principal component asset index is 0.050. We rescale this coefficient by multiplying it by $\frac{\rho'\delta}{\sigma_{pc}}$ to make it comparable to the estimates where the assets are entered separately. In this formula ρ is the bivariate correlations between each asset and school enrollment, δ are the scoring factors, and σ_{pc} is a vector of the standard deviations of the asset variables. The adjusted coefficient on the asset index is 0.098, over 40 percent 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 any of the *post hoc* estimators that use only six of the assets.

6 Conclusion

We have proposed a new estimator for the case where the 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 *post hoc* 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 estimators compared in this paper. The need to estimate ρ, the covariances between the proxies and the unobserved factor, introduces additional noise into the estimates that is not present in an *ad hoc* index variable. The Monte Carlo simulation presented in Section 4 suggest that this source of variance may not be particularly large and is outweighed by the large reduction in bias in the estimates themselves. 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," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. II, Elsevier, 1984, pp. 1321–1393.

Blau, David M., "The Effect of Income on Child Development," *The Review of Economics and Statistics*, 1999, *81* (2), 261–276.

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

Case, Anne, Darren Lubotsky, and Christina Paxson, "Economic Status and Health in Childhood: The Origins of the Gradient," *American Economic Review*, forthcoming.

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*, February 2001, *38* (1), 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*, 2000, *115*, 811–846.

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

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

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*, 1975, *70* (351), 631–639.

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

Mauro, Paolo, "Corruption and Growth," *Quarterly Journal of Economics*, 1995, *110* (3), 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," Working paper 2001–18, Federal Reserve Bank of Chicago 2001.

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

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

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

A Proofs

Lemma A.1

1. $|A + aa'| = |A|\left(1 + a'A^{-1}a\right)$, *if* A *is nonsingular.*
2. $(A + ab')^{-1} = A^{-1} - \frac{A^{-1}ab'A^{-1}}{1+b'A^{-1}a}$, *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

$$\text{plim}\widehat{\beta} \quad = \quad \frac{\beta}{\delta'\rho}\left(1 - \frac{\delta'\Sigma_{UU}\delta}{\sigma_x^2\left(\delta'\rho\right)^2 + \delta'\Sigma_{UU}\delta}\right)$$

i.e.

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

By contrast

$$\text{plim}b \quad = \quad \beta\sigma_x^2\left(\Sigma_{XX}\right)^{-1}\rho$$

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

$$\Sigma_{XX}^{-1}\rho \quad = \quad \frac{\Sigma_{UU}^{-1}\rho}{1 + \sigma_x^2\rho'\Sigma_{UU}^{-1}\rho} \tag{17}$$

Hence

$$\text{plim}\widehat{\rho}'b = \beta\left(1 - \frac{1}{1 + \sigma_x^2\rho'\Sigma_{UU}^{-1}\rho}\right)$$

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

$$\frac{1}{1 + \sigma_x^2\rho'\Sigma_{UU}^{-1}\rho} = \frac{|\Sigma_{UU}|}{|\Sigma_{XX}|} \tag{18}$$

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

$$\rho'\Sigma_{UU}^{-1}\rho \geq \frac{\left(\delta'\rho\right)^2}{\delta'\Sigma_{UU}\delta}$$

for any non-zero choices of ρ and δ.

Since Σ_{UU} is a non-singular covariance matrix, by the spectral theorem for symmetric matrices it can be decomposed as

$$\Sigma_{UU} = PDP^{'}$$

where P is an orthogonal matrix of eigenvectors

$$P = \left[\begin{array}{cccc} p_1 & p_2 & \ldots & p_k \end{array} \right]$$

and $D = diag\left(\lambda_1, \ldots, \lambda_k\right)$ is the matrix of eigenvalues, with $\lambda_i > 0$, $\forall i$. This is equivalent to writing

$$\Sigma_{UU} = \lambda_1 p_1 p_1^{'} + \ldots + \lambda_k p_k p_k^{'}$$

and it follows that

$$\Sigma_{UU}^{-1} = \frac{1}{\lambda_1} p_1 p_1^{'} + \ldots + \frac{1}{\lambda_k} p_k p_k^{'}$$

Hence

$$\begin{aligned} \rho^{'} \Sigma_{UU}^{-1} \rho &= \sum_i \frac{1}{\lambda_i} \left(p_i^{'} \rho \right)^2 \\ \delta^{'} \Sigma_{UU} \delta &= \sum_i \lambda_i \left(p_i^{'} \delta \right)^2 \end{aligned}$$

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

$$w = P^{'} \rho, \; v = P^{'} \delta$$

Note that

$$\delta^{'} \rho = v^{'} w$$

(since P is orthogonal), i.e.

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

Equality holds only if

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

for some real number c, i.e.

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

■

Proof of proposition 2.

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

and by using equations 17 and 18

$$e_i^{'} \left(\Sigma_{XX}\right)^{-1} \rho \sigma_x^2 = \frac{|\Sigma_{UU}|}{|\Sigma_{XX}|} \left(\sigma_x^2 e_i^{'} \Sigma_{UU}^{-1} \rho \right)$$

Now

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

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

$$e_i^{'} \Sigma_{UU}^{-1} = \frac{1}{|\Sigma_{UU}|} \begin{bmatrix} S_{i1} & S_{i2} & \ldots & S_{ik} \end{bmatrix}$$

i.e.

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

The term $\sum_j S_{ij} \rho_j$, however, is identical to the value of the determinant if row i of matrix Σ_{UU} were replaced by $\rho^{'}$, i.e.

$$\sum_j S_{ij} \rho_j = \left| \Sigma_{UU}^{\rho(i)} \right|$$

Consequently

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

■

Table 1: Monte Carlo Simulation Results

	Post hoc estimator			*Average of*	
Number of	*true*	*estimated*	*Average of*	*standardized*	*First principal*
proxies	p'*s*	p'*s*	*proxies*	*proxies*	*component*
1	66.77	66.77	66.77	66.77	66.77
	(4.69)	(4.69)	(4.69)	(4.69)	(4.69)
2	73.14	73.52	62.62	63.71	63.45
	(7.26)	(7.88)	(16.00)	(16.19)	(16.65)
3	77.67	78.37	66.26	66.12	69.72
	(7.39)	(8.40)	(14.96)	(15.01)	(15.10)
4	80.95	81.85	69.57	68.74	73.48
	(6.91)	(8.27)	(13.70)	(13.73)	(12.98)
5	83.37	84.40	71.97	70.58	75.94
	(6.18)	(8.06)	(12.12)	(12.36)	(11.14)
10	89.58	91.10	77.56	75.57	81.78
	(4.21)	(7.29)	(8.37)	(8.68)	(7.09)
15	92.53	94.18	80.16	78.00	84.23
	(3.26)	(7.21)	(6.71)	(7.13)	(5.50)
20	94.26	96.00	81.46	79.36	85.47
	(2.76)	(7.18)	(5.99)	(6.35)	(4.89)
MSE	[40.57]	[67.55]	[379.75]	[466.25]	[235.08]

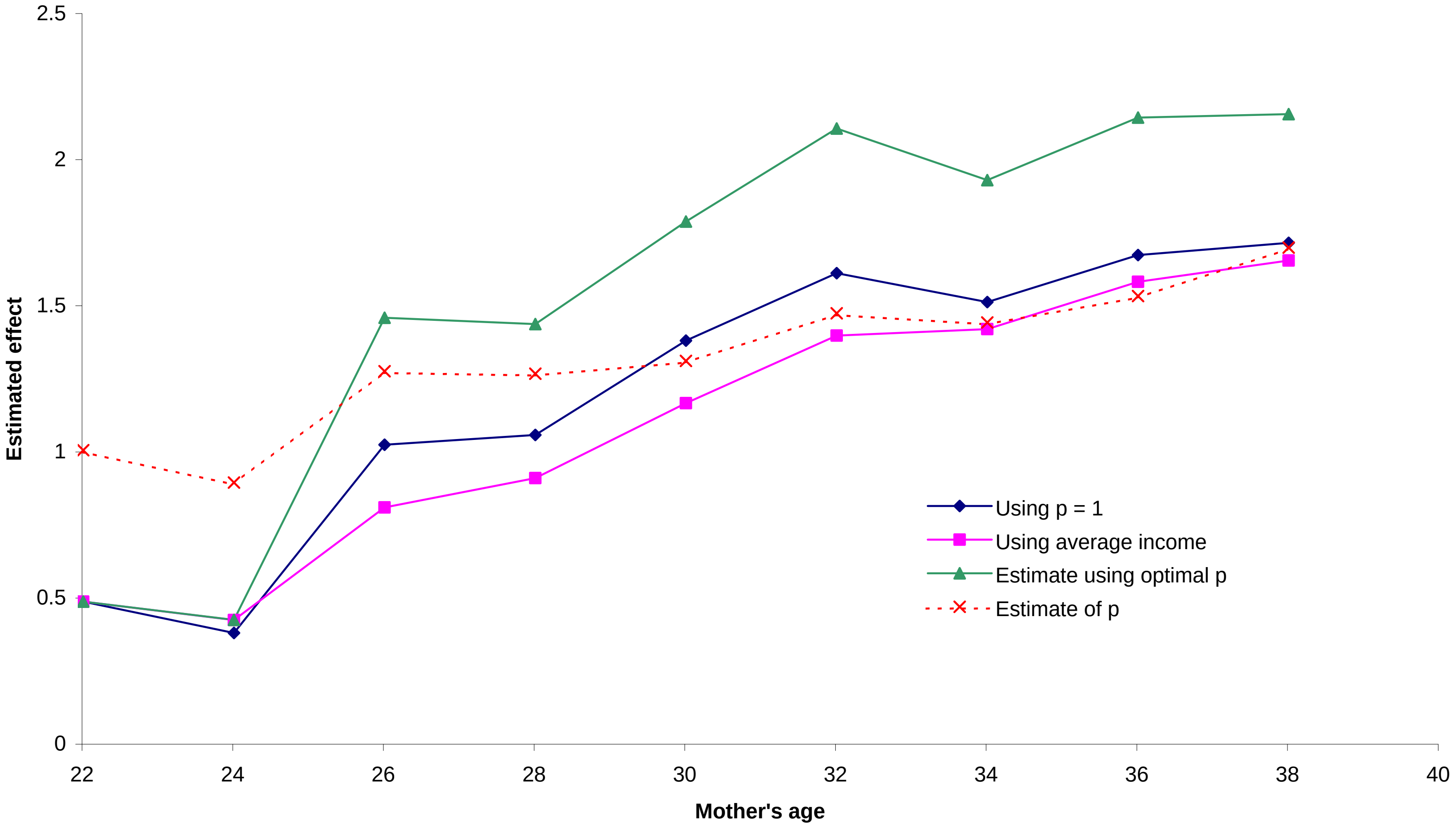
Note: The table shows the mean and standard deviation of each estimator. The mean square error for the models with 20 proxies are given in the last line. See text for details.

Table 2: Measuring the Effect of Wealth on Children's School Attendance In India

				Proxy set				Principal components
	Rho	1	2	3	4	5	6	relative weights
# Rooms in house	1.000	0.009	0.011		0.013			1.000
		(0.001)	(0.001)		(0.001)			
Refrigerator	0.121	-0.035	-0.035		-0.029			0.546
		(0.006)	(0.006)		(0.005)			
Clock or watch	0.452	0.089	0.123		0.132			0.429
		(0.003)	(0.003)		(0.003)			
Type of toilet								
Flush	0.275	0.046	0.074			0.113		0.077
		(0.004)	(0.004)			(0.004)		
Latrine	0.091	0.052	0.069			0.094		1.007
		(0.004)	(0.004)			(0.004)		
Sewing machine	0.291	0.039	0.067			0.098		0.922
		(0.003)	(0.003)			(0.003)		
VCR	0.038	-0.014	-0.033			-0.025		0.662
		(0.008)	(0.008)			(0.008)		
Radio	0.353	0.034	0.054				0.084	1.249
		(0.003)	(0.003)				(0.003)	
Drinking water from								
Pump	-0.226	0.002	-0.032				-0.028	-0.154
		(0.011)	(0.004)				(0.004)	
Open source	-0.001	0.023	0.010				0.025	-0.694
		(0.012)	(0.005)				(0.005)	
Cooking fuel is wood/ dung/coal	-0.243	0.004	-0.008				-0.017	-1.097
		(0.004)	(0.004)				(0.004)	
Television	0.332	0.005		0.034	0.038			0.962
		(0.004)		(0.004)	(0.004)			
Non-drinking water from								
Pump	-0.203	-0.014		-0.023	-0.038			-0.283
		(0.011)		(0.004)	(0.004)			
Open source	-0.028	0.021		0.029	0.007			1.099
		(0.012)		(0.005)	(0.005)			
Fan	0.414	0.022		0.053	0.097			1.216
		(0.004)		(0.004)	(0.003)			
Bicycle	0.188	0.019		0.036		0.046		0.934
		(0.003)		(0.003)		(0.003)		
Car	0.019	-0.062		-0.064		-0.062		-0.839
		(0.010)		(0.010)		(0.010)		
Kitchen in separate room	0.311	0.064		0.085		0.097		0.666
		(0.003)		(0.003)		(0.003)		
Motorcycle	0.129	-0.004		0.000			0.013	1.055
		(0.005)		(0.005)			(0.005)	
Electric lighting	0.447	0.114		0.137			0.162	0.616
		(0.003)		(0.003)			(0.003)	
Livestock	-0.179	-0.007		-0.009			-0.006	-0.573
		(0.003)		(0.003)			(0.003)	
Number of proxies		18	9	9	6	6	6	18
Estimated "wealth" effect		0.170	0.136	0.132	0.129	0.105	0.116	0.098

Note: Data is 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.

Figure 1: The Effect of Family Income on Children's Reading Comprehension Score



Note: Data is 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 *p* s in order to be on the same scale as the other two estimators.