Testing for a common latent variable in a linear regression

Wittenberg, Martin

School of Economics, University of Cape Town

31 March 2007

Testing for a common latent variable in a linear regression*

Martin Wittenberg

School of Economics

University of Cape Town

This version March 2007

Abstract

We present a test of the hypothesis that a subset of the regressors are all proxying for the same latent variable. This issue will be of interest in cases where there are several correlated measures of elusive concepts such as misgovernance or corruption; in analyses where key variables such as income are not measured at all and one is forced to rely on various proxies; and where the key regressors are badly measured and one is trying to extract a stronger signal from the regression by adding additional proxies as suggested by Lubotsky and Wittenberg (2006).

We apply this test in three contexts, each characterised by a different estimation challenge arising from data limitations. We reexamine Mauro's (1995) use of three institutional quality measures in his study of corruption and growth. Here several variables, each potentially measured with error, may all be proxies for a single factor: the quality of governance. Our test suggests that the latent variable is driven primarily by the "red tape" measure, rather than the "corruption" variable on which Mauro focuses.

*I thank Rob Garlick, Murray Leibbrandt, Duncan Thomas, Chris Udry and participants at an African Econometric Society Conference for useful comments on this paper. I also thank Rob Garlick for sterling research assistance.

Secondly, we look at the correlates of body mass among black South African women. The key variable of interest, namely "wealth" is not measured at all. Consequently we construct an index from a series of asset variables as suggested by Filmer and Pritchett (2001). Our test shows that some assets have independent impacts on the dependent variable. Once this is recognised the "asset index" comes apart.

Finally we analyse the determinants of sleep among young South Africans. The income variable in the survey is badly measured and we supplement it with asset proxies. The test again suggests that some assets are not proxying for the badly measured income variable. We can nevertheless get a substantially stronger signal on the income variable.

Keywords: measurement error, proxy variables, specification test, asset index

JEL codes: C12, C13, C52

1 Introduction

In many applied contexts the underlying theory is not strong enough to pin down the precise form of a regression model. Finding a reasonable specification in these circumstances is likely to be a somewhat haphazard enterprise. One common problem is that there may be several candidate measures of the core concepts of the theory. In other cases these may not be measured at all, or measured extremely badly, and the researcher is forced to rely on proxy variables. The theory may also be vague on the role of potential covariates.

The textbook treatment of specification error provides an initial assessment of the impact of these problems. In the most benign case, that of adding an irrelevant regressor, misspecification leads to a loss of efficiency. In other cases, such as the omission of a relevant variable or choice of incorrect functional form, the result of misspecification is inconsistency of the estimates. Given these asymmetric costs, researchers are tempted to include additional covariates to better isolate the impact of the variables of interest. If these additional regressors do not belong in the "structural" relationship, their coefficients should be statistically insignificant. Nevertheless in a recent article Lubotsky and Wittenberg (2006) suggest that it is possible for multiple proxies of a single "structural" variable to be significant in a

multiple regression if they are all subject to measurement error. Indeed, they suggest that deliberately misspecifying the regression by including all available proxies will, in general, be preferable to creating a summary variable from the available proxies. They provide a method by which a lower bound on the structural coefficient can be estimated from the misspecified regression. We summarise the main points of their paper in the appendix. The validity of the Lubotsky-Wittenberg (LW) estimation procedure hinges, of course, on the question whether it is plausible that these regressors are, indeed, all functioning as proxies for the same latent variable.

In this paper we develop a test of the hypothesis that several regressors are acting as proxies for a single latent variable. This question is likely to be interesting in several different contexts:

1.1 The role of institutions in economic performance

The importance of institutions for economic performance has become increasingly recognised. Indeed, institutions seem to make a significant difference in cross-country growth regressions (Acemoglu, Johnson and Robinson 2001, Bosworth and Collins 2003). Nevertheless many of the institutional measures are correlated with each other, so that it is difficult to know whether a particular variable is important in itself or is proxying for some other factor (Fedderke and Klitgaard 1998). Indeed, Bosworth and Collins (2003) note that a variable as important as education becomes insignificant in a regression context once one includes a measure of institutional quality.

Furthermore in many cases there are multiple plausible indicators of the particular institution that is being measured. In these cases authors have to plump for one or another, or find some way of combining them. In a frequently cited study, Mauro (1995) had a choice of three different indicators of corruption and misgovernance. He chose to average his indicators. Fundamental to this procedure is the judgement that these indicators are all proxying for the same underlying latent variable. It may, however, be the case that these institutional quality variables are themselves only proxying for the effect of some other variable, e.g. the level of education of the population or the size of the middle class.

In this paper we will re-examine Mauro's conclusions. Our test confirms that these three observables capture the same underlying variable and that they have a separate impact in the regression. However, the latent variable is driven primarily by the "red tape" measure, rather than the "corruption" variable on which Mauro focuses.

1.2 Estimation of "wealth effects" through asset variables

There are many data sets in which income or expenditure data is missing, but information is available on household assets. In an influential paper, Filmer and Pritchett (2001) suggest that the first principal component of the asset variables can be used as a proxy for household wealth. This method has led to a small growth industry in the analysis of the Demographic and Health Surveys and, indeed, many other data sets.

In applying this method one would like to know whether the assets do, in fact, proxy for a common latent variable. Is it possible that any of the assets have an independent effect on the outcome variable? If so, this would contaminate the interpretation of the index. It is, after all, a linear combination of the proxies.

We show below that this concern is well-founded. We examine the relationship between the Body Mass Index and such an asset index in the case of the South African Demographic and Health Survey. Obesity is becoming recognised as an important health risk among black South African women, with 22% of all these women being obese. Understanding the economic and social processes underpinning this trend would be important. The asset index is strongly positively correlated with the body mass of black South African women. Nevertheless several of the assets do not proxy only for income. Ownership of a car (unsurprisingly) has a strong positive impact on body mass. Similarly television ownership has a direct impact. Indeed after all the assets that do **not** proxy only for "wealth" are stripped out, the remaining "wealth" effect is reduced by three-quarters and is significant only at the 10% level on a one-tailed test.

1.3 Improving on a badly measured income variable

Generally speaking asset variables have been used in cases where expenditure or income data is completely unavailable. The Lubotsky and Wittenberg (2006) (LW) estimation procedure can be used, however, to correct for a badly measured income variable also. It would stand to reason that augmenting the information from income with that obtained from the asset variables should improve the estimate. Indeed Szalontai (2006) uses the LW procedure in precisely this way and shows that the coefficient on the badly measured income variable almost doubles. As our comments on the asset proxies should highlight, this procedure is valid only if the assets do not have independent effects.

We reexamine Szalontai's example, the relationship between sleep and income in South Africa's time use survey. Instead of focussing on adults, however, we examine the relationship among children. We find the same marked increase in the size of the income coefficient. Nevertheless we also find that access to electricity seems to have an independent impact on sleep times. Ignoring this would lead to an overly high estimate by the LW procedure.

In all three cases having access to a test for a common latent variable changes the empirical analysis. We wish to emphasize that the three cases represent different uses of the procedure:

- In the first case there are many correlated measures each of which could plausibly feature in the structural equation. In the absence of the test any decisions about how to summarise the variables, which ones to include and which to leave out, is somewhat ad-hoc.

- In the second case the concept of interest (income or wealth) is well-defined but is represented by a summary measure extracted from a set of proxies. Our test examines whether the main regression is correctly specified or whether some of the proxies belong in it, in which case the coefficient on the "asset index" is misleading.

- The final example deals with a case in which the regression equation is **deliberately** misspecified in order to counteract attenuation bias on a badly measured variable which is included in the main regression. The test examines whether the LW procedure is

valid or not. In the process of developing this test we also introduce a more efficient version of the LW estimator.

The plan of the discussion is as follows. In the next section we introduce the central insight of this paper: that the estimator of ρ, the coefficient on the latent variable in the LW framework, can best be thought of as a particular type of instrumental variable estimator. If there are covariates in the model it transpires that more instruments may become available. This in turn opens up the possibility of testing the validity of the moment conditions. We introduce this in section 3. First we show how the proxies can be tested individually and then, in sections 4 and 5 we do so for the system as a whole. This test of the overidentifying restrictions can be thought of as an omnibus test for all sorts of failures of the model. Acceptance of the null hypothesis, i.e. of model validity, is equivalent to accepting that the variables tested are all proxying for a common latent variable.

In section 6 we demonstrate the procedures by means of the three empirical examples. Section 7 concludes.

2 The regression with proxy variables and one covariate

Let us consider the model

$$y = x\beta + z\gamma + \varepsilon$$

where we assume that z is measured accurately and $cov\,(x, z) = \sigma_{xz}$ is non-zero. We assume that we have two proxy variables

$$\begin{aligned} x_0 &= x + u_0 \\ x_1 &= x\rho_1 + u_1 \end{aligned}$$

where we assume that the error variables u_j are uncorrelated with x, z and ε. The estimated model is

$$y = x_0 b_0 + x_1 b_1 + z\theta + \eta$$

We have selected a different symbol for the coefficient of z to indicate that typically this will not be (not even asymptotically) equal to γ.

As it stands there are no intercepts in this model. We have implicitly projected all variables on a constant, i.e. we are writing the model in deviations form. This means that all the errors will automatically have zero mean, so that $E\left(u_j u_j'\right) = \sigma_j^2 I_n$. Hence we can write the model either in terms of covariances or the expected values of products.

The empirical information at our disposal is summarised in the following correlation matrix:

$$\begin{bmatrix} \beta^2\sigma_x^2 + 2\beta\gamma\sigma_{xz} + \gamma^2\sigma_z^2 + \sigma_\varepsilon^2 & \beta\sigma_x^2 + \gamma\sigma_{xz} & \beta\rho_1\sigma_x^2 + \gamma\rho_1\sigma_{xz} & \beta\sigma_{xz} + \gamma\sigma_z^2 \\ & \sigma_x^2 + \sigma_0^2 & \rho_1\sigma_x^2 + \sigma_{01} & \sigma_{xz} \\ & & \rho_1^2\sigma_x^2 + \sigma_1^2 & \rho_1\sigma_{xz} \\ & & & \sigma_z^2 \end{bmatrix}$$

We observe that ρ_1 is now **overdetermined**: $\rho_1 = cov\left(y, x_1\right)/cov\left(y, x_0\right)$ and $\rho_1 = cov\left(z, x_1\right)/cov\left(z, x_0\right)$ This raises the question as to how to estimate ρ_1 most efficiently.

Consideration of the form of the GMM estimates of ρ_1 given above, suggest that it can be thought of as an instrumental variables estimator in the regression of x_1 on x_0. We have

$$\begin{aligned} x_1 &= x\rho_1 + u_1 \\ &= \left(x_0 - u_0\right)\rho_1 + u_1 \\ &= x_0\rho_1 + \upsilon_1 \end{aligned}$$

Note that this cannot be estimated consistently by OLS, since x_0 is correlated with u_0 and hence υ_1. Since we have assumed that $cov\left(y, x_1\right) \neq 0$, and $cov\left(y, \upsilon_1\right) = 0$ (i.e. neither of the measurement error terms are correlated with any of the terms in the main regression), y is a legitimate instrument for x_1 in this regression. This may seem somewhat surprising given that y is actually the dependent variable in a regression in which x_1 is an explanatory variable. In addition note that if $cov\left(z, x_1\right) \neq 0$ and z is uncorrelated with any of the error terms, then z is also a legitimate instrument. This suggests that the optimal estimation strategy for ρ_1 is to use two-stage least squares, with y and z as instruments for x_0.

What happens if z is correlated with any of the u_j terms? In that case it would clearly be invalid to use z as an instrument. Note, however, that we can write

$$x_1 = x_0\rho_1 + z\phi_1 + v_1 \tag{1}$$

where by construction v_1 is now uncorrelated with z. We can estimate ρ_1 and ϕ_1 in the standard way, using y as an instrument for x_0 and z as an instrument for itself. The estimate of ρ_1 obtained in this way is numerically identical to the "covariate adjusted" estimator suggested in Lubotsky and Wittenberg (2006), except that we do not have to obtain the residuals first.

One advantage of writing the proxy in the form of equation 1 is that it is possible to use the estimate of ϕ_1 to calculate a LW estimate of γ similar to the LW estimate of β. In this case the proxy variable adjusted estimate would be

$$\widehat{\gamma} = \widehat{\theta} + \widehat{\phi}_1\widehat{b}_1$$

Although z is not mismeasured, the fact that b_1 is asymptotically nonzero due to the measurement error in x requires this LW adjustment to be made. With the adjustment, however, the overall bias in the estimate of γ would be lower than in any regression with some other linear function of x_0 and x_1, in particular a regression in which only x_0 is used as an explanatory variable. This follows, by extension, from the results in Lubotsky and Wittenberg (2006).

3 Estimation and testing of ρ and ϕ proxy by proxy

We will now consider the more general model

$$\begin{aligned}
y &= x\beta + Z\gamma + \varepsilon && \text{(2a)}\\
x_0 &= x + u_0 && \text{(2b)}\\
x_1 &= x\rho_1 + Z_{(1)}\phi_1 + u_1 && \text{(2c)}\\
&\cdots\\
x_k &= x\rho_k + Z_{(k)}\phi_k + u_k && \text{(2d)}
\end{aligned}$$

where $cov(\varepsilon, u_j) = 0$. Furthermore we assume that $E(Z'u_0) = \mathbf{0}$. Since Z is assumed to contain a column of ones this implies that u_0 has mean zero. Consequently x_0 is an instance of "classical measurement error". If this assumption is violated, the LW procedure can still be applied, but the coefficients of the Z variables that are correlated with u_0 will not be correctable by means of the procedure outlined below. We allow the other k proxies for x, viz. $x_1, \ldots, x_k$ to be more flexibly defined. In particular we suppose that the deviations from the latent variable x may be systematic and explicable in terms of some of the covariates.

We assume that the matrix of covariates Z can be partitioned as

$$Z = \left[\begin{array}{cc} Z_{(i)} & Z_{-(i)} \end{array} \right]$$

into variables $Z_{(i)}$ that have a direct effect on the proxy x_i, when controlling for the latent variable x, as well as variables $Z_{-(i)}$ that are correlated with x_i only through x. If there is a variable that is orthogonal to x and x_i we will include it in $Z_{(i)}$ to remove any ambiguity. It will generally be the case that $Z_{(i)}$ will contain at least a column of ones for the intercept. With constants in all equations we can assume that all error terms are mean zero. Finally we assume that $Z_{-(i)}$ is also non-empty. Note that in many applications including the ones reported on below we may be able to partition the matrix in the same way for every proxy.

With these assumptions, we can write the i-th proxy as

$$\begin{aligned} x_i &= x_0\rho_i + Z_{(i)}\phi_i + u_i - u_0\rho_i \\ &= M_{(i)}\delta_i + \upsilon_i \end{aligned} \tag{3}$$

where $M_{(i)} = \left[\begin{array}{cc} x_0 & Z_{(i)} \end{array} \right]$ and $\delta_i = \left[\begin{array}{c} \rho_i \\ \phi_i \end{array} \right]$. Furthermore by our assumptions,

$$Z'\left(x_i - x_0\rho_i - Z_{(i)}\phi_i\right) = 0 \tag{4a}$$

$$y'\left(x_i - x_0\rho_i - Z_{(i)}\phi_i\right) = 0 \tag{4b}$$

and hence ρ_i and the coefficient vector ϕ_i can be consistently estimated by instrumental variables. In addition, provided that $Z_{-(i)}$ is non-empty our estimates are over-determined.

An appropriate estimator for this equation would therefore be the generalised IV estimator, or two-stage least squares estimator

$$\widehat{\delta}_i = \left(M'_{(i)} Z_y \left(Z'_y Z_y \right)^{-1} Z'_y M_{(i)} \right)^{-1} M'_{(i)} Z_y \left(Z'_y Z_y \right)^{-1} Z'_y x_i \tag{5}$$

where Z_y is the matrix of instruments, i.e.

$$Z_y = \left[\begin{array}{cc} y & Z \end{array} \right] \tag{6}$$

With these estimates for ρ_i and ϕ_i the LW estimates of β and γ will be given by

$$\widehat{\beta} = b_0 + \sum_{i=1}^{k} \widehat{\rho}_i b_i \tag{7a}$$

$$\widehat{\gamma}_j = \widehat{\theta}_j + \sum_{i=1}^{k} \widehat{\phi}_{ij} b_i \tag{7b}$$

where $\widehat{\theta}_j$ is the unadjusted coefficient on z_j in the multiple regression with all the proxies included and where $\widehat{\phi}_{ij} = 0$ if $z_j \notin Z_{(i)}$.

If $Z_{-(i)}$ is non-empty it is possible to test for the validity of the LW model by means of a test of the overidentifying restrictions. A particularly easy form of such a test is described in Davidson and MacKinnon (1993, p.236). They show that n times the uncentered R^2 from a regression of the IV residuals on the instruments Z is distributed as χ^2 with degrees of freedom equal to the degree of overidentification. In this case the degree of overidentification is exactly equal to the number of variables in $Z_{-(i)}$.

The null hypothesis for this test is that the moment conditions in equations 4 are all valid. This hypothesis could be rejected for a number of reasons:

1. ε is correlated with any of the u_i terms
2. one or more of the covariates in $Z_{-(i)}$ is correlated with any of the u_i terms
3. The proxy model is misspecified, i.e. it is not the case that

$$x_i = x_0 \rho_i + Z_{(i)} \phi_i + \upsilon_i$$

This could be due to either the fact that it is not the case that

$$x_i = x\rho_i + Z_{(i)}\phi_i + u_i$$

or that it is not the case that

$$x_0 = x + u_0$$

4. The main regression model is misspecified

Most of these would be reasons for being sceptical about the validity of applying the LW model. If the test fails for the second reason, however, it would be possible to re-partition the covariates and include the offending variables in the controls $Z_{(i)}$. If all covariates end up being included it is, of course, no longer possible to test the model. This may also be an indication that the LW model is of dubious value.

As an aside, we note that the "covariates adjusted" estimator of ρ suggested in Lubotsky and Wittenberg (2006) is identical to the estimate obtained if $Z_{(i)} = Z$, i.e. if none of the covariates was a legitimate instrument. We would expect therefore that the procedure outlined above should lead to more efficient estimates than that given in the original paper.

4 Systems estimation of ρ and ϕ

According to the model given in equations 2 there are k equations of the type 3 to estimate. We can "stack" these equations in the standard way:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} M_{(1)} & 0 & \cdots & 0 \\ 0 & M_{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_{(k)} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_k \end{bmatrix} + \begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \vdots \\ \upsilon_k \end{bmatrix} \tag{8}$$

$$\mathbf{x}_v = \mathbf{M}\boldsymbol{\delta} + \boldsymbol{\upsilon}_v \tag{9}$$

In the special case where the same control variables $Z_{(i)} = Z_u$ are used in every equation the matrix $\mathbf{M}$ takes on the particularly simple form

$$\begin{aligned} \mathbf{M} &= I_k \otimes \begin{bmatrix} x_0 & Z_u \end{bmatrix} \\ &= I_k \otimes M_x \end{aligned} \tag{10}$$

where $M_x = \left[\begin{array}{cc} x_0 & Z_u \end{array} \right]$.

The systems version of the proxy by proxy estimation outlined above would be to define the matrix of instruments

$$\begin{aligned} \mathbf{Z}_v &= \left[\begin{array}{cccc} Z_y & 0 & \cdots & 0 \\ 0 & Z_y & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_y \end{array} \right] \\ &= I_k \otimes Z_y \end{aligned} \tag{11}$$

and then define the "2SLS" estimator

$$\widehat{\boldsymbol{\delta}}_{2SLS} = \left(\mathbf{M}' \mathbf{Z}_v \left(\mathbf{Z}_v' \mathbf{Z}_v \right)^{-1} \mathbf{Z}_v' \mathbf{M} \right)^{-1} \mathbf{M}' \mathbf{Z}_v \left(\mathbf{Z}_v' \mathbf{Z}_v \right)^{-1} \mathbf{Z}_v' \mathbf{x}_v \tag{12}$$

which is numerically equal to the proxy by proxy estimation of $\boldsymbol{\delta}$ given in equation 5.

This estimator, however, may not be efficient, since the covariance matrix of $\boldsymbol{\upsilon}_v$ is not diagonal. It is obvious that the vector υ_i and the vector υ_j are correlated. In fact a typical covariance between an element of υ_i and υ_j will be given by $cov\left(\upsilon_i, \upsilon_j\right) = cov\left(u_i - u_0\rho_i, u_j - u_0\rho_j\right) = \sigma_{ij} - \rho_j\sigma_{0i} - \rho_i\sigma_{0j} + \rho_i\rho_j\sigma_0^2$. Let us denote this as υ_{ij}. Then

$$\begin{aligned} E\left(\boldsymbol{\upsilon}_v \boldsymbol{\upsilon}_v'\right) &= \Psi \\ &= \Sigma \otimes I_n \end{aligned}$$

where

$$\Sigma = \left[\begin{array}{cccc} \upsilon_{11} & \upsilon_{12} & \cdots & \upsilon_{1k} \\ \upsilon_{12} & \upsilon_{22} & \cdots & \upsilon_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{1k} & \upsilon_{2k} & \cdots & \upsilon_{kk} \end{array} \right]$$

This suggests that the system of proxy equations could be more efficiently estimated as a system, taking these cross-equation correlations into account.

A brief consideration of the formal properties of the model shows that the GLS estimator of the system is identical to the three stage least squares estimator described in the literature

(Davidson and MacKinnon 1993, Mittelhammer, Judge and Miller 2000). This means that estimation of $\boldsymbol{\delta}$ and hence ρ can proceed with readily available software. Furthermore the asymptotic properties of this estimator are well established.

Nevertheless it is useful to briefly go through the derivation of the estimator, since we will use some of the intermediate results in section 5. We follow the treatment in Mittelhammer et al. (2000, pp.462–465) and develop the estimator as the optimal GMM estimator from the population moment condition

$$E\left(\mathbf{Z}_v'\left(\mathbf{x}_v - \mathbf{M}\boldsymbol{\delta}\right)\right) = \mathbf{0} \tag{13}$$

This yields the sample counterpart

$$\frac{1}{n}\mathbf{Z}_v'\left(\mathbf{x}_v - \mathbf{M}\boldsymbol{\delta}\right) = \mathbf{0}$$

and hence the GMM estimator

$$\widehat{\boldsymbol{\delta}}\left(\mathbf{W}\right) = \left[\mathbf{M}'\mathbf{Z}_v\mathbf{W}\mathbf{Z}_v'\mathbf{M}\right]^{-1}\mathbf{M}'\mathbf{Z}_v\mathbf{W}\mathbf{Z}_v'\mathbf{x}_v \tag{14}$$

for some positive definite weighting matrix $\mathbf{W}$. Indeed with $\mathbf{W} = \left(\mathbf{Z}_v'\mathbf{Z}_v\right)^{-1}$ we get the "2SLS" estimator of equation 12. The **optimal** weighting matrix, however, is

$$\begin{aligned} w_*^{-1} &= cov\left(n^{-\frac{1}{2}}\mathbf{Z}_v'\boldsymbol{v}_v\right) \\ &= n^{-1}E\left(\mathbf{Z}_v'\boldsymbol{v}_v\boldsymbol{v}_v'\mathbf{Z}_v\right) \\ &= n^{-1}E\left(\left(I_k \otimes Z_y\right)'\left(\Sigma \otimes I_n\right)\left(I_k \otimes Z_y\right)\right) \\ &= \Sigma \otimes E\left(n^{-1}Z_y'Z_y\right) \end{aligned} \tag{15}$$

$\mathbf{W}$ is therefore proportional to $\Sigma^{-1} \otimes \left(Z_y'Z_y\right)^{-1}$ and the corresponding GMM estimator is given by

$$\boldsymbol{\delta}_{GMM} = \left[\mathbf{M}'\left(\Sigma^{-1} \otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)\mathbf{M}\right]^{-1}\mathbf{M}'\left(\Sigma^{-1} \otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)\mathbf{x}_v \tag{16}$$

The matrix Σ is unknown, but can be consistently estimated using the residuals from any consistent GMM estimator. The 2SLS estimator (12) is particularly convenient in this regard. We have

$$\widehat{\Sigma}_{ij} = \frac{1}{n}\widehat{v}_i'\widehat{v}_j$$

where $\widehat{v}_i$ and $\widehat{v}_j$ are the vectors of residuals from the i-th and j-th proxy estimation respectively. Using this in place of Σ gives the estimated weighting matrix

$$\widehat{\mathbf{W}} = \widehat{\Sigma}^{-1} \otimes \left(Z_y' Z_y\right)^{-1} \tag{17}$$

and correspondingly the estimated optimal GMM estimator $\widehat{\boldsymbol{\delta}}_{GMM}$ or three-stage least squares estimator.

In the case where the same control variables $Z_{(i)} = Z_u$ are used in every equation, we can substitute equation 10 into equation 16. This yields

$$\begin{aligned} \widehat{\boldsymbol{\delta}}_{GMM} &= \left[\widehat{\Sigma}^{-1} \otimes M_x' Z_y \left(Z_y' Z_y\right)^{-1} Z_y' M_x\right]^{-1} \left[\widehat{\Sigma}^{-1} \otimes M_x' Z_y \left(Z_y' Z_y\right)^{-1} Z_y'\right] \mathbf{x}_v \\ &= \left\{I_k \otimes \left(M_x' Z_y \left(Z_y' Z_y\right)^{-1} Z_y' M_x\right)^{-1} M_x' Z_y \left(Z_y' Z_y\right)^{-1} Z_y'\right\} \mathbf{x}_v \end{aligned} \tag{18}$$

This, however, is numerically identical to the proxy by proxy estimation of $\boldsymbol{\delta}$ by 2SLS! This finding parallels the result that least squares estimation and GLS estimation of a SUR system with **identical** X matrices are equal (see for instance Mittelhammer et al. 2000, p.453).

5 Specification testing

Just as it is possible to test the overidentifying restrictions proxy by proxy, it is possible to do so on the system as a whole. The appropriate test statistic in this case will be given by

$$\left[\mathbf{Z}_v' \left(\mathbf{x}_v - \mathbf{M}\widehat{\boldsymbol{\delta}}_{GMM}\right)\right]' \widehat{\mathbf{W}} \left[\mathbf{Z}_v' \left(\mathbf{x}_v - \mathbf{M}\widehat{\boldsymbol{\delta}}_{GMM}\right)\right] \xrightarrow{d} \chi^2(d)$$

(Mittelhammer et al. 2000, pp.438–9) where $\widehat{\mathbf{W}}$ is given by equation 17 and d is the degree of overidentification, i.e. the number of instruments used in the system as a whole minus the number of moment equations. The test statistic τ can be simplified to

$$\widehat{v}_v' \left(\widehat{\Sigma}^{-1} \otimes Z_y \left(Z_y' Z_y\right)^{-1} Z_y'\right) \widehat{v}_v$$

where $\widehat{v}_v$ is the vector of stacked IV residuals. Now note that this is equivalent to

$$\widehat{v}_v' \left(I_k \otimes Z_y \left(Z_y' Z_y\right)^{-1} Z_y'\right)' \left(\widehat{\Sigma}^{-1} \otimes I_n\right) \left(I_k \otimes Z_y \left(Z_y' Z_y\right)^{-1} Z_y'\right) \widehat{v}_v$$

Let

$$\widehat{\widehat{v}} = \left(I_k \otimes Z_y \left(Z_y' Z_y \right)^{-1} Z_y' \right) \widehat{v}_v$$

This is the vector of fitted values from the artificial regression of the residuals $\widehat{v}_v$ on the matrix of instruments $\mathbf{Z}_v$. Hence the test statistic can also be written as

$$\widehat{\widehat{v}}' \left(\widehat{\Sigma}^{-1} \otimes I_n \right) \widehat{\widehat{v}}$$

In the trivial case where this estimator is applied to a "system" of only one equation $\widehat{\Sigma}^{-1} = \widehat{\sigma}^{-2}$ and the test statistic is identical to nR_u^2 in the regression of $\widehat{v}_v$ on $\mathbf{Z}_v$. Whenever $\widehat{\Sigma}^{-1} \otimes I_n$ is **not** equal to $\widehat{\sigma}^{-2} I_{kn}$, however, this statistic will not be computable in this manner. Instead, let

$$\widehat{\Sigma}^{-1} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1k} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \cdots & \alpha_{kk} \end{bmatrix}$$

and let $\widehat{\widehat{v}}_i$ refer (in the obvious way) to the vector of fitted values corresponding to the i-th proxy then

$$\begin{aligned} \widehat{\widehat{v}}' \left(\widehat{\Sigma}^{-1} \otimes I_n \right) \widehat{\widehat{v}} &= \begin{bmatrix} \widehat{\widehat{v}}_1' & \widehat{\widehat{v}}_2' & \cdots & \widehat{\widehat{v}}_k' \end{bmatrix} \begin{bmatrix} \alpha_{11} I_n & \alpha_{12} I_n & \cdots & \alpha_{1k} I_n \\ \alpha_{21} I_n & \alpha_{22} I_n & \cdots & \alpha_{2k} I_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1} I_n & \alpha_{k2} I_n & \cdots & \alpha_{kk} I_n \end{bmatrix} \begin{bmatrix} \widehat{\widehat{v}}_1 \\ \widehat{\widehat{v}}_2 \\ \vdots \\ \widehat{\widehat{v}}_k \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^k \alpha_{i1} \widehat{\widehat{v}}_i' & \sum_{i=1}^k \alpha_{i2} \widehat{\widehat{v}}_i' & \cdots & \sum_{i=1}^k \alpha_{ik} \widehat{\widehat{v}}_i' \end{bmatrix} \begin{bmatrix} \widehat{\widehat{v}}_1 \\ \widehat{\widehat{v}}_2 \\ \vdots \\ \widehat{\widehat{v}}_k \end{bmatrix} \\ &= \sum_{i=1}^k \sum_{j=1}^k \alpha_{ij} \widehat{\widehat{v}}_i' \widehat{\widehat{v}}_j \end{aligned} \tag{19}$$

If we define the matrix B as

$$B_{ij} = \widehat{\widehat{v}}_i' \widehat{\widehat{v}}_j$$

and we note that B is symmetric, it is easy to see that

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_{ij}\widehat{\widehat{\upsilon}}_i'\widehat{\widehat{\upsilon}}_j = tr\left(\widehat{\Sigma}^{-1}B\right)$$

If the artificial regression fits perfectly, so that $\widehat{\widehat{\upsilon}}_j = \widehat{\upsilon}_j$, then $B = n\widehat{\Sigma}$ and the test statistic will be kn.

The important point to note is that the calculation of the test statistic does not require the formation of any $kn \times kn$ matrices. Indeed provided that one uses the same partition of the Z matrix in each proxy equation there is no need to compute anything other than the proxy-by-proxy two stage least squares estimates and the residuals from them. Even in this case one may want to utilise three stage least squares routines since these conveniently calculate the matrix $\widehat{\Sigma}$. Certainly one will need to do so if one wanted to test restrictions on the $\boldsymbol{\rho}$ vector.

In the case where the same controls are used in each proxy equation it is easy to calculate the degrees of freedom for the hypothesis test. If we let the number of variables that are **not** in Z_u be l, i.e.

$$l = rank\left(Z\right) - rank\left(Z_u\right)$$

then the degrees of freedom for the test will be $d = lk$, i.e. the number of instruments $\left(rank\left(Z\right)+1\right)k$ used less the number of moment equations, i.e. $\left(1+rank\left(Z_u\right)\right)k$.

The null hypothesis for this test is that the moment condition given in equation 13 holds. As we noted above, a rejection of the null can occur for a number of different reasons:

- A correlation between ε and any of the u_i terms. In this case y would cease to be an appropriate instrument. This implies that the proxy variable x_i is not proxying only for the latent variable x, but should be in the main regression.

- A correlation between any of the z_j variables used as instruments and any of the u_i terms. In this case z_j should be treated as a control in equation i and not as an instrument.

- A misspecification of any of the proxy variable equations

- A misspecification of the main regression

In short rejection of the null hypothesis should probably be taken as evidence that the model should be seriously rethought, although it may be technically possible to respecify it in ways that provide consistent estimates of ρ.

In our empirical applications we view rejection of the model as a sign that the model is inadequate. For instance, in the "corruption" example we will see that the system test rejects the assumption that the three "institutional" variables are proxying for secondary school enrolment.

In the case of the "asset index" example we have to be somewhat more circumspect, because it is clear by considering the nature of the variables involved that some of the other covariates should be treated as controls and not as instruments. Nevertheless even once that correlation is acknowledged the system test still rejects the idea that all the assets are simply proxying for a commmon latent variable ("wealth"). We interpret this as evidence that some of the assets belong in the main regression.

6 Applications: proxying for institutions, wealth and income

In order to get a sense how these techniques perform on real data we apply them to three contexts where we suspect that there may be common latent variables:

1. A cross-country study of the relationship between institutions and investment.
2. A study on a Demographic and Health Survey, where there is no income or expenditure data, but there are asset proxies
3. A study on a data set where the income variable is measured very badly, but there are asset proxies

The three examples present some of the spectrum of cases where our specification test may prove useful. The first is an instance of a regression where the underlying concepts

are elusive and difficult to capture. As a result different measures may be available and the analyst needs to decide whether it is permissible to combine these in some way.

In the second case the variable of interest is not measured at all, but we have a number of proxies available. The problem is, that some of these may have effects independent of the latent variable they are supposedly proxying for.

The last example is, perhaps, the cleanest case for the Lubotsky and Wittenberg (2006) procedure, since there is a ready made proxy to serve as numeraire – the income variable itself. Here the key issue is whether one can get a measure of the impact of that variable that is subject to less attenuation bias.

6.1 The link between corruption and investment

Mauro's study of the link between corruption and growth has been the landmark study in this area. He suggests that a key link in the chain between corruption and growth is investment – corruption and bad institutions have the effect of significantly lowering the investment rate.

In Table 1 we reexamine the link between corruption and investment using a data set that is as close as possible to the original. The first two columns of that table reproduce directly the results from the first two columns of Mauro's Table VI (1995, p.699). We re-run those regressions with our data set in columns three and four. The coefficients are close to those reported by Mauro and the significance levels are close as well. The main result is that the "bureaucratic efficiency" and corruption index both have a direct and statistically highly significant impact on the level of investment (relative to GDP). We note that high values of the index correspond to good institutions, so that lower corruption and better efficiency are associated with higher investment.

In column five we insert all three of the indexes that Mauro considered. Only the "red tape" index turns out to be significant, although the other two indexes also record positive coefficients. The reason for this is may simply be that the indexes are all too highly correlated with each other and subject to too much measurement error. Indeed Mauro's "bureaucratic efficiency" index is simply the arithmetic average of the three indexes and is constructed with the express purpose of reducing measurement error.

We test the hypothesis that all three variables are simply proxying for a common latent variable in column six. We use the "corruption index" as the chief proxy x_0, i.e the proxy that fixes the scale of the latent variable. As the test statistic and p-value at the foot of that column indicate, the data accept the hypothesis that there is a common latent variable. We see that the optimal linear combination of the indexes leads to a sizable increase in the measured coefficient from 0.015 to 0.021, which is around a 40% improvement.

We can calculate the LW index that would correspond to this estimate. It is given by

$$LW = 0.11 \text{ corruption} + 0.66 \text{ red tape } + 0.35 \text{ judicial effectiveness}$$

The coefficients do not sum to one, which suggests that although all three variables are theoretically scored on a scale from 1 to 10, the "red tape" and "judicial effectiveness" scores are effectively operating on a more compressed scale than the corruption ones. Indeed the range and standard deviation of those two variables is somewhat smaller than those of the corruption index. The main point is that the latent variable that is picked by the data looks much more like a "red tape" variable than a "corruption" one.

Column seven checks to see whether the "bureaucratic effectiveness" index can be improved on. In this regression we are including the bureaucratic effectiveness index as proxy x_0 and also include the "red tape" and "judicial effectiveness" index. In this case the improvement brought about by the LW procedure is minor. In short simply averaging the three indexes goes almost all the way to reducing the attenuation bias in this case. When compared to the optimal LW index the data, however, again suggest a major role for red tape. The index in this case is given by

$$LW = 0.29 \text{ bureaucratic efficiency } + 0.51 \text{ red tape } + 0.22 \text{ judicial effectiveness}$$

Since the bureaucratic efficiency index is the simple average of the corruption, red tape and judicial effectiveness indices, the data again suggest that the optimal weighting of the three indexes is around 10% of the corruption index, 60% of the red tape one and 30% of the judicial effectiveness one.

In the last two columns of table 1 we consider whether the three indexes of good governance are merely proxying either for the level of income in the society (as measured by GDP

per capita) or education. As the χ^2 statistics and p-values indicate, the data soundly reject these hypotheses.

In summary the analyses support Mauro's decision to combine the three indexes. They even suggest that averaging them is close to optimal. However the data also make clear that most of the work in the regression is being done by the "red tape" index. If there is a latent variable that all three indexes are proxying for, then it looks much more like a "red tape" one than a corruption one.

6.2 Body mass and asset ownership

There has been a rapid increase in obesity, even in developing countries (Popkin 1999). South Africans have followed this trend, with many showing high body mass even in otherwise poverty stricken communities (Case and Deaton 2005). This has led to an increase in the prevalence of hypertension and strokes in contexts where one might not have expected to see this (Kahn and Tollman 1999). Understanding the socio-economic correlates of obesity is therefore of considerable interest.

The main data set available for exploring this relationship is the South African Demographic and Health Survey. Like other such surveys, it has extensive health information but very few socio-economic variables. In particular it has no income or expenditure information. A standard procedure now in such cases is to estimate an asset index by principal components (Filmer and Pritchett 2001) and to use this as a proxy for wealth.

In Table 2 we show a set of regressions in which the body mass of black South African women is regressed on asset proxies and a set of individual and household attributes. We focus on women because obesity is particularly prevalent in this group. The dependent variable is the body mass index (BMI) defined as weight (in kilogram) divided by height (in metres) squared. The first column shows that a one unit increase (equivalent to a standard deviation increase) in the asset index is associated with an average increase of 0.599 in the BMI. This translates into an increase of 1.4kg (3 lbs) for a woman of average height (1.577m). This suggests that women of higher economic status would also be heavier, which runs counter to the trend in many developed countries where obesity is a problem

disproportionately of low income communities. This finding is reinforced by the strongly positive coefficient on education.

In column 2 we include the proxies separately while column 3 provides an estimate on the coefficient of the underlying latent variable if we go on the assumption that all the assets are proxying for a common latent variable. We have used telephone ownership as our x_0 proxy. Since we would really like an estimate of the impact of expenditure, we projected telephone ownership on log household expenditure in the South African Income and Expenditure survey. This led to the estimate

$$tel = -2.175 + 0.2585 \ln (\text{expenditure})$$

i.e. $\widehat{\rho}_0 = 0.2585$. By rescaling the telephone variable by the reciprocal of 0.2585 it should be on a scale where a one unit increase in the latent variable will correspond to a one unit increase in $\log (\text{expenditure})$. As it is, the standard deviation of the log of household expenditure happens to be 1.076. The coefficient on the LW index in column 3 should therefore correspond roughly to a standard deviation increase in log household expenditure. It should therefore be directly comparable to the coefficient on the asset index. As the theory would suggest, the LW procedure produces a coefficient that is markedly larger.

Nevertheless before jumping to conclusions it would be prudent to test that the asset variables are, indeed, all proxying for a common latent variable. In order for this test to pass muster we need to allow for correlations between the "employment" and "education" variables and the assets. These two variables almost certainly have effects on the accumulation of assets independent of total expenditure (or wealth). Employment is likely to feature because many household assets (such as television sets) are purchased through loan agreements and employment makes such acquisitions easier, even controlling for household wealth. Education is likely to influence asset accumulation both as a taste shifter and as a necessary input into the utilisation of some assets. It therefore does not make sense to think of asset ownership being correlated with these two variables only through the latent variable. The specification test is thus implemented using age, the square of age, household composition (number of children and number of adults) and being a smoker as exogenous variables.

Strikingly, the specification test roundly rejects this model. The individual specification tests on six of the assets are particularly vehemently rejected, these being access to electricity, television, refrigerator, bicycle, car and sheep/cattle ownership. Once these are allowed to have independent effects (as in column 4 of Table 2), the specification test is happy to accept that the remaining assets could proxy for a common latent variable. Interestingly not all of the assets rejected by the specification test turn out to have statistically signficant coefficients. Whatever they might be measuring, however, it is not the same kind of thing as what the other assets are capturing. So, for instance, it is quite likely that ownership of sheep and cattle represents not just wealth, but traditional values and a rural location. To the extent to which rural diets and lifestyles are lagging behind the shifts that are driving the obesity epidemic in urban areas, this index may function differently in the regression. Indeed we observe that the point estimates in columns 2 and 4 are negative.

Two of the assets stand out for their very strong effects. Car ownership is associated with almost a full unit increase in mean BMI. This represents a 2.4kg (5.3lb) increase for a woman of average height. Television has an identically strong effect. The specification test rejects the idea that these are merely income, expenditure or wealth effects. They are lifestyle effects which come with the acquisition of those particular assets.

When we put these variables into a multiple regression together with the asset index (in column 5) we notice that the coefficients are 22% smaller. In the case of car ownership the variable is significant only at the 10% level. We would therefore make misleading inferences. Of course one is asking for trouble if one includes assets twice over – in the asset index and separately in the regression. In column six we therefore recalculate the asset index only over variables that we do not intend to include separately in the regression. This regression provides a very similar picture to that given by the LW procedure (in column 4), except that the coefficient on the asset index is 30% smaller. It is completely insignificant, while the LW asset index would be significant at the 12% level.

Nevertheless neither regression makes a strong case for an independent and sizable wealth or income effect. It looks as though improvements in socio-economic status lead to increases in body mass largely through the more sedentary lifestyle that they help to buy. A simple uncritical use of the asset index obscures this mechanism.

6.3 Sleep and income

In the previous cases the latent variable was not measured at all. We now consider an example where the variable is measured, but badly. The application is the demand for sleep among school going South Africans. The pioneering study on the economic determinants of sleep was by Biddle and Hamermesh (1990). They argued that the length of sleep was not purely biologically determined but responded to economic signals. The opportunity cost of sleep is the wage foregone and so one would expect high wage earners to sleep less than low wage ones. They show this with US time use data. Szalontai (2006) examines the same relationship on South African data and finds similar results. Because his income variable is badly measured he employs the Lubotsky-Wittenberg procedure using a number of asset variables to extract a stronger signal. Indeed the coefficient increases by 75% in absolute value.

We will re-examine this procedure but on a different subsample, viz. school age children. A recent study suggested that sleep in this group also strongly decreases with income (Wittenberg 2005). In this case the opportunity cost of sleep cannot be the wage, so there must be other explanations, including the entertainment opportunities foregone.

The data for the South African studies comes from the South African time use survey conducted by Statistics South Africa (Budlender, Chobokoane and Mpetsheni 2001). One of the problems with investigating economic relationships on this data set is that the income information is very poor. In this it is not unique: in all surveys there is some trade off between the breadth of issues covered and the quality of information obtained on particular variables. Surveys which ask a simple income question and do not extensively probe are unlikely to get quality income data. The extent of the problem can be seen in Table 3, which contrasts the total household income distribution from the Time Use Survey and the expenditure and income distributions from an Income and Expenditure Survey conducted by Statistics South Africa in the same year, i.e. 2000.

It is clear that total household income in the Time Use Survey has been measured with considerable error. It also seems clear that this error is not classical measurement error in the form given in equation 2b. There seems to be a systematic underreporting of incomes,

so that the measurement error term u_0 has a non-zero mean. This means that the intercept term in the regression will have an additional unknown bias. Furthermore it is plausible that the scale of the measurement error may be correlated with some of the covariates. The LW "corrected" coefficients of these terms should therefore be viewed with some caution.

The scale of the underreporting however also increases the attractiveness of the LW procedure. It seems clear that households with very low reported incomes, but possessing fridges, stoves and televisions must be better off than households without those. Including asset proxies would control for income much better than using the reported income categories alone.

For the purposes of the regressions we have turned the discrete income categories into a continuous variable, by taking the midpoints of the categories and twice the lower bound of the open category. The latter is recommended by Charles Simkins (personal communication), based on the fact that the income distribution at the top end is roughly Pareto with parameter just under two. This adds an additional level of noise to the variable which will, however, be of secondary influence compared to the underreporting shown above.

In column 1 of Table 4 we report a simple regression of minutes spent sleeping during an average night during the school week (Monday to Friday). The sample is restricted only to individuals who are actually observed attending school in the time diaries. The coefficient on log income is around -10, which would translate to a 35 minute difference between children with the median income (R600) and the richest children (R20 000). Adding in the proxies we see a marked drop in this coefficient, but significant coefficients on "fridge" and "TV". The LW estimates in columns (3) and (4) make different assumptions about which covariates are correlated with the proxies and which are independent of the measurement errors.

In column (3) we assume that all covariates are exogenous. This is effectively what Szalontai assumed also. The coefficient is 85% larger, but we observe that the specification test roundly rejects the underlying assumptions. Given South Africa's history it is unlikely that the racial dummies will be exogenous. Black South Africans are likely to have an "asset deficit" which is not simply a function of income. We might assume also that some of the household composition variables might affect the acquisition of particular types of assets. In column (4) we therefore treat only the attributes of the child and the timing of the

survey as exogenous. The individual proxy specification tests now come out as accepting the null hypothesis, except in the case of television ownership. We have consequently allowed television ownership to have an independent effect on sleeping times. The LW estimates given in column (4) still show a marked increase on the income coefficient compared to column (1). The point estimate on income corresponds to about 48 minutes difference in sleep times between the median income and the top. This is a nontrivial impact, particularly if one bears in mind the additional 23 minutes loss of sleep attendant on owning a television!

As this regression performs well on all the specification tests we might be happy to leave the issue. There seem to be two factors that matter: income raises the opportunity costs of sleep, while ownership of a television does so even when controlling for income. This simple picture is, however, muddied by the results shown in the remaining columns. In column five we have added in two "infrastructure" proxies - use of electricity for lighting and living in a brick dwelling. The LW estimates reported in column (6) correspond to the same model as that accepted by the data in column (4), except that these additional proxies have been used in the LW procedure. In this case, however, the specification test roundly rejects the validity of the model. More particularly it suggests that electricity should be in the main model. It also raises the prospect that perhaps it is not television *per se* that is important for sleep, but access to electricity and the attendant opportunities.

In columns six and seven we explore two competing hypotheses:

- In column six we have the hypothesis that only the log of income and television ownership belong in the structural equation. Access to electricity is interpreted in the LW model as just another proxy for mismeasured income. As we have noted above, this model is not supported by the specification tests - either for the "electricity" proxy equation or for the system as a whole.

- In column seven we explore the hypothesis that only the log of income and access to electricity belong in the structural equation. TV ownership is relegated to the status of a proxy for income. This model is accepted by the data at conventional levels of significance despite the fact that the TV ownership proxy has a statistically significant and large coefficient in the regression.

The procedure that we have employed in adjudicating whether television or electricity are structural is reminiscent of Granger causality tests. Like those tests it is possible for the answers to be less clear cut. The data might have suggested that both variables belong in the main regression or that neither do. Of course tests are hardly ever completely decisive and the power of the tests may be inadequate.

Reviewing the evidence available in Table 4 we come to the conclusion that a point estimate of around -18 is probably a reasonable guess at the impact of income on sleep. The naive LW procedure did not lead to a distorted estimate in this case, although it did not reveal the additional impact of electricity or TV. The point estimate would correspond to a difference of about an hour sleep per day between kids at the median income versus children in the top income bracket. In addition to this children with electrified houses would sleep 24 minutes a day less. One of the opportunities opened up by larger incomes is, of course, the ability to watch television. So television probably matters - but as an outcome rather than a structural determinant. Indeed an initial look at patterns of television watching suggests that television viewing may be related to income in inverse U fashion - increasing with income and then decreasing at the top. The richest children probably have many more non-TV opportunities (e.g. internet chat rooms) to while away the nights. In short there are non-technical reasons for believing that the specification tests have picked out a reasonable model.

7 Conclusion

In all three empirical applications the ability to test for a common latent variable has added to our understanding of the underlying relationships. In the case of the institutional quality scores we could show that it was reasonable to try to combine them. We could also pinpoint which of the variables were doing most of the work in the regression. This suggested that the impact being estimated was that of "red tape" more than "corruption".

In the second example our test really picked the composite index apart. It showed that some of its components, notably car and television ownership, were having large and independent effects on body mass. This suggested that the wealth effect was being mediated

by lifestyle changes.

In the third case the main issue was to estimate the coefficient on log income more reliably. We showed that a naive application of the Lubotsky-Wittenberg procedure (2006) would miss the fact that some assets seem to have impacts independent of income. In particular, access to electricity seems to open up activities (such as television watching) that crowd out sleep. In this case we also showed that the tests can be used to think about the "structural" relationship between proxies. We did this by means of a Granger-style "proxy test". We tested whether:

- TV could be proxying for income controlling for the presence of electricity or
- Electricity could be proxying for income controlling for the presence of TV.

Ultimately, however, these techniques are no substitute for thinking about the underlying relationships theoretically. Indeed the LW procedure cannot in any real way substitute for proper judgement or "fix" bad data. If income is badly measured the LW procedure provides the promise that one can get an additional measure of people's control of resources by taking account of their assets. But this is second prize compared to better quality income information. Furthermore as these examples should make clear the "fix" needs to be run with proper caution.

In this paper we have shown how the LW procedure can be run more efficiently. The empirical results suggest that it can significantly strengthen the estimated coefficients. Under certain circumstances this could make a difference to how one interprets the output. Nevertheless, before applying it, it would be prudent to run the specification test developed in this paper.

A Appendix: The LW procedure

Lubotsky and Wittenberg (2006) analyse the model

$$\begin{aligned} y &= \beta x^* + \varepsilon \\ x_j &= \rho_j x^* + u_j \end{aligned}$$

where $\rho_1 = 1$ and the errors u_j are uncorrelated with ε and x^* and have mean zero, but are otherwise unrestricted. They show that ρ_j can be consistently estimated by

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

They suggest that in general one should include **all** proxies in the main regression and aggregate the coefficients up according to the formula:

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

where b is the vector of OLS regression coefficients (i.e. b_j is the coefficient on the j-th proxy in that regression). They argue that this will produce a lower bound on the true coefficient β. Furthermore this lower bound will have less attenuation bias than estimates obtained by any other linear combination of the proxy variables. They also show that this procedure implicitly constructs an index from the proxies, which can be explicitly calculated as

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

If there are covariates in the model, they suggest that the procedure be run after the effect of the covariates has been removed by projecting both y and the proxies on the covariates.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson, "The colonial origins of comparative development: an empirical investigation," *American Economic Review*, 2001, *91* (5), 1369–1401.

Biddle, Jeff E. and Daniel S. Hamermesh, "Sleep and the Allocation of Time," *Journal of Political Economy*, 1990, *98* (5 Part 1), 922–943.

Bosworth, Barry P. and Susan M. Collins, "The empirics of growth: An update," *Brookings Papers on Economic Activity*, 2003, (2), 113–179.

Budlender, Debbie, Ntebaleng Chobokoane, and Yandiswa Mpetsheni, *A Survey of Time Use: How South African women and men spend their time*, Pretoria: Statistics South Africa, 2001.

Case, Anne and Angus Deaton, "Health and wealth among the poor: India and South Africa compared," *American Economic Review Papers and Proceedings*, 2005.

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

Fedderke, Johannes and Robert Klitgaard, "Economic Growth and Social Indicators: An Exploratory Analysis," *Economic Development and Cultural Change*, 1998, *46* (3), 455–489.

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.

Heston, A., R. Summers, and B. Aten, "Penn World Table version 6.2," September 2006. Center for International Comparisons of Production, Income and Prices, University of Pennsylvania.

Kahn, K. and S.M. Tollman, "Stroke in rural South Africa — contributing to the little known about a big problem," *South African Journal of Medicine*, 1999, *89* (1), 63–65.

Lubotsky, Darren and Martin Wittenberg, "Interpretation of regressions with multiple proxies," *Review of Economics and Statistics*, 2006, *88* (3), 549–562.

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

Mittelhammer, Ron C., George G. Judge, and Douglas J. Miller, *Econometric Foundations*, Cambridge: CUP, 2000.

Popkin, Barry M., "Urbanization, Life Style Changes and the Nutrition Transition," *World Development*, 1999, *27* (11), 1905–1916.

Summers, R. and A. Heston, "A New Set of International Comparisons of Real Product and Price Levels Estimates for 130 Countries, 1950 – 1985," *Review of Income and Wealth*, 1988, *34*, 1–25.

Szalontai, Gabor, "The demand for sleep: A South African Study," *Economic Modelling*, 2006, *23*, 854–874.

Wittenberg, Martin, "How young South Africans spend their time," *Loisir et Société/Society and Leisure*, 2005, *28* (2), 635–652.

Table 1 Investment and corruption

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Original Mauro 1	Original Mauro 2	Rerun Mauro 1	Rerun Mauro2	With all indexes	Corruption proxy	Bureaucratic eff proxy	GDP proxy	Education proxy
GDP	-0.008	-0.006	-0.010	-0.007	-0.011	-0.011	-0.011	0.013	-0.011
	(0.006)	(0.007)	(0.006)+	(0.007)	(0.006)+	(0.006)+	(0.006)+	(0.008)+	(0.006)+
Education	0.060	0.111	0.018	0.075	0.010	0.010	0.010	0.010	0.224
	(0.062)	(0.066)+	(0.054)	(0.060)	(0.057)	(0.054)	(0.054)	(0.049)	(0.067)**
population growth	-1.373	-0.620	-1.514	-0.803	-1.804	-1.804	-1.805	-1.804	-1.804
	(0.995)	(1.02)	(0.892)+	(0.996)	(0.885)*	(0.873)*	(0.873)*	(0.752)*	(0.847)*
bureaucratic efficiency	0.019		0.023				0.024		
	(0.005)**		(0.005)**				(0.006)**		
corruption		0.013		0.015	0.002	0.021			
		(0.004)**		(0.005)**	(0.005)	(0.005)**			
red tape					0.014				
					(0.006)*				
judicial efficiency					0.008				
					(0.006)				
Constant	0.104	0.114	0.103	0.123	0.109	0.109	0.109	0.109	0.109
	(0.034)**	(0.036)**	(0.034)**	(0.037)**	(0.036)**	(0.035)**	(0.035)**	(0.032)**	(0.034)**
Observations			57	57	57	57	57	57	57
R-squared	0.51	0.47	0.50	0.43	0.51	0.51	0.51	0.51	0.51
System test: Chi2						8.7	8.7	22.9	18.2
df						6	6	6	6
P value:						0.189	0.193	0.001	0.006

Dependent variable: Investment/GDP. Robust standard errors in parentheses. Standard errors on the LW proxy coefficient calculated by bootstrap with 1000 replications. Standard errors on the original Mauro regressions calculated from the published t-statistics.

+ significant at 10%; * significant at 5%; ** significant at 1%

Sources: Data on aggregate investment/GDP averaged from 1960 to 1985, GDP per capita in 1960 and the proportion of the population enrolled in secondary education in 1960 are obtained from the Penn World Table version 4 (Summers and Heston, 1988). The index for population growth is calculated from the Penn World Table version 6.2 by taking an average of the annualized growth rates between 1960 and 1985 (Heston, *et al*, 2006). Summary statistics agree very closely with those reported by Mauro.

Table 2 The relationship between BMI and assets among African women

	(1)	(2)	(3)	(4)	(5)	(6)
	PC1	All proxies	LW1	LW2	PC1	PC2
proxy	0.566		0.804	0.160	0.250	0.112
	(0.063)**		(0.108)**	(0.102)	(0.176)	(0.096)
employed	0.167	0.226	0.275	0.218	0.208	0.213
	(0.241)	(0.242)	(0.248)	(0.242)	(0.241)	(0.241)
education	0.125	0.118	0.133	0.118	0.120	0.121
	(0.030)**	(0.030)**	(0.032)**	(0.032)**	(0.030)**	(0.030)**
age	0.596	0.596	0.596	0.596	0.596	0.596
	(0.036)**	(0.036)**	(0.038)**	(0.036)**	(0.036)**	(0.036)**
age^2	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005
	(0.000)**	(0.000)**	(0.000)**	(0.000)**	(0.000)**	(0.000)**
children	0.141	0.150	0.150	0.150	0.146	0.145
	(0.052)**	(0.053)**	(0.053)*	(0.054)*	(0.053)**	(0.053)**
adults	-0.119	-0.125	-0.125	-0.125	-0.119	-0.116
	(0.065)+	(0.066)+	(0.067)+	(0.064)	(0.066)+	(0.065)+
smoker	-2.309	-2.270	-2.270	-2.270	-2.299	-2.305
	(0.381)**	(0.382)**	(0.422)**	(0.412)**	(0.382)**	(0.382)**
telephone		0.127				
		(0.083)				
electricity		0.548		0.541	0.363	0.572
		(0.257)*		(0.249)+	(0.303)	(0.256)*
television		0.898		0.960	0.751	0.972
		(0.259)**		(0.256)**	(0.308)*	(0.255)**
refrigerator		0.421		0.409	0.218	0.467
		(0.285)		(0.291)	(0.355)	(0.282)+
bicycle		-0.372		-0.378	-0.548	-0.407
		(0.315)		(0.322)	(0.334)	(0.313)
car		1.017		0.971	0.754	1.005
		(0.330)**		(0.333)**	(0.403)+	(0.328)**
sheep/cattle		-0.121		-0.069	0.015	-0.087
		(0.287)		(0.262)	(0.296)	(0.285)
radio		0.434				

		(0.259)+				
personal computer		0.456				
		(0.981)				
washing machine		-0.292				
		(0.472)				
motorcycle		-0.460				
		(1.457)				
constant	13.063	11.710	11.710	11.710	12.401	12.012
	(0.901)**	(0.903)**	(0.926)**	(0.887)**	(0.965)**	(0.896)**
Observations	4342	4342	4342	4342	4342	4342
R-squared	0.12	0.12	0.12	0.12	0.12	0.12
System test: Chi2			367.7	23.2		
df			50	20		
P value:			0.000	0.278		

Dependent variable: Body Mass Index. Standard errors in parentheses. Bootstrapped standard errors are given in columns (3) and (4)
+ significant at 10%; * significant at 5%; ** significant at 1%
PC1: Principal components Assets Index calculated over all 11 assets
PC2: Principal components Assets Index calculated over the 5 assets not used in the regression (telephone, radio, personal computer, washing machine and motorcycle).
LW1: LW index calculated over all 11 assets. Telephone ownership used to calibrate the index. Correlation between “employed” and “education” variables and proxies allowed for.
LW2: LW index calculated over 5 assets. Telephone ownership used to calibrate the index. Correlation between “employed”, “education”, the other six asset variables and the proxies allowed for.
Sources: Own calculations from the South African Demographic and Health Survey (1998)

Table 3 Comparing Income distributions in the IES and Time Use Survey

Category (per month)	Income and Expenditure Survey 2000		Time Use Survey 2000
	Expenditure	Income	Income
	(proportions)		
R 0-399	0.076	0.097	0.204
R 400-799	0.195	0.215	0.286
R 800-1199	0.160	0.144	0.141
R 1200-1799	0.151	0.134	0.103
R 1800-2499	0.105	0.092	0.068
R 2500-4999	0.151	0.149	0.093
R 5000-9999	0.089	0.090	0.074
R 10000+	0.073	0.079	0.031
	1.000	1.000	1.000

Sources: Own calculations from the South African Income and Expenditure Survey, 2000 and the Time Use Survey, 2000.

Table 4 The relationship between sleep, income and assets among schoolgoing adolescents in South Africa

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
N=1867	Income only	Assets	LW1	LW2	Assets plus infrastructure	LW3	LW4
loginc	-9.98	-4.05	-18.49	-13.65	-4.22	-13.43	-19.62
	(2.14) ***	(2.36) +	(4.12)***	(4.16)**	(2.36) +	(4.41)**	(4.58)***
Household size	-0.06	0.91	0.91	1.06	1.01	1.04	1.39
	(1.34)	(1.33)	(1.33)	(1.56)	(1.33)	(1.46)	(1.48)
Number of children	-0.17	-1.36	-1.36	-1.29	-1.61	-1.29	-1.89
	(1.89)	(1.88)	(1.88)	(2.07)	(1.88)	(2.02)	(2.09)
Coloured	-2.74	1.91	1.91	1.86	1.27	1.87	4.05
	(8.41)	(8.66)	(8.66)	(10.07)	(8.65)	(10.88)	(10.51)
Indian	13.18	16.31	16.31	21.11	18.6	21.35	29.45
	(12.89)	(13.04)	(13.04)	(12.53)+	(13.05)	(11.30)+	(12.51)*
White	13.71	19.02	19.02	26.42	19.74+	26.22	37.23
	(9.75)	(11.4) +	(11.4) +	(13.20)*	(11.38)	(12.89)*	(13.16)**
age	-3.9	-4.83	-4.83	-4.5	-4.67	-4.67	-4.67
	(1.1) ***	(1.1) ***	(1.1) ***	(1.1) ***	(1.1) ***	(1.1) ***	(1.1) ***
Years of education	-5.13	-3.98	-3.98	-4.49	-4.21	-4.21	-4.21
	(1.21) ***	(1.21) **	(1.21) **	(1.2) ***	(1.21) **	(1.21) **	(1.21) **
gender	-4.74	-3.42	-3.42	-3.85	-3.54	-3.54	-3.54
	(3.6)	(3.56)	(3.56)	(3.56)	(3.55)	(3.55)	(3.55)
TV		-18.02		-23.48	-16.3	-23.52	
		(4.59) ***		(4.63)***	(4.63) ***	(4.56)***	
Washing machine		-1.35			-2.04		
		(7.59)			(7.58)		
vacuum		-4.16			-4.57		
		(8.47)			(8.45)		
fridge		-10.98			-7.78		

		(4.95) *			(5.08)		
phone		-3			-2.04		
		(4.54)			(4.57)		
stove		-7.38			-2.92		
		(4.73)			(4.95)		
radio		-5.9			-6.22		
		(4.66)			(4.66)		
car		-0.45			-1.15		
		(4.93)			(4.94)		
clock		-1.18			-1.06		
		(4.56)			(4.56)		
electric lights					-15.5		-23.17
					(5.06) **		(4.67)***
brick dwelling					0.96		
					(4.13)		
Controls for stratum and province	Y	Y	Y	Y	Y	Y	Y
Controls for tranche and day of week	Y	Y	Y	Y	Y	Y	Y
System test: Chi2			909.83	83.64		117.86	100.91
df			225	72		90	90
p-value			0.000	0.164		0.026	0.203

Notes

Standard errors are given in parentheses and are corrected for clustering. Standard errors for the LW estimates were calculated by means of a clustered bootstrap with 200 replications. Significance level: + 10% * 5% ** 1% *** 0.1%

LW1: All covariates deemed exogenous

LW2, LW3: Age, education, gender of child as well as tranche and day of week deemed exogenous. TV deemed to have independent effect on sleep.

LW4: Age, education, gender of child as well as tranche and day of week deemed exogenous. Electricification deemed to have independent effect on sleep.

Sources: Own calculations from the South African Time Use Survey, 2000.