DataFirst Technical Papers

Weighing the value of Asset Proxies: The case of the Body Mass Index in South Africa

by
Martin Wittenberg

Technical Paper Series
Number 13

About the Author(s) and Acknowledgments

I thank Susan Godlonton for excellent research assistance. Some parts of this paper were presented at seminars at SALDRU, Princeton and Michigan. I thank all participants for useful comments.

This is a joint DataFirst and SALDRU paper as part of the Mellon Data Quality Project

Recommended citation

Wittenberg, M. (2009). Weighing the value of Asset Proxies: The case of the Body Mass Index in South Africa. A DataFirst Technical Paper Number 13. Cape Town: DataFirst, University of Cape Town

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

Weighing the value of Asset Proxies: The case of the Body Mass Index in South Africa

Martin Wittenberg*
School of Economics and SALDRU
University of Cape Town

This version January 2009

1 Introduction

There are many household surveys, e.g. the Demographic and Health Surveys, that carry a wealth of useful information but in which information of interest to economists, chiefly incomes, is extremely badly measured or missing altogether. In many of these surveys, however, there are questions about asset ownership. These might be used to proxy for income. Indeed some authors have suggested that asset-based measures of well-being may even be better than income or expenditure-based ones, since they may reflect the long-run welfare of the household better. They may also be more accurately measured (Filmer and Pritchett 2001, Sahn and Stifel 2003).

The key question that needs to be confronted is how one should aggregate up the very different types of assets. Since the seminal paper by Filmer and Pritchett (2001) one of the most popular ways of dealing with this issue has been to construct asset indices based on the first principal component of the available asset variables. This procedure provides that linear combination of the asset variables which explains the greatest proportion of their joint variance. It is natural to think of this as extracting an index of affluence – hence the title of the paper "Estimating wealth effects without expenditure data – or tears". While principal components is now undoubtedly the most popular way of dealing with the asset variables, there are a number of alternatives, such as simply adding up the number of assets. A number of reviews have examined some of these alternatives (Bollen, Glanville and Stecklov 2002, Houweling, Kunst and Mackenbach 2003, Howe, Hargreaves and Huttly 2008, Montgomery, Gragnolati, Burke and Paredes 2000).

The ease with which principal components based indices can be constructed has led to the proliferation of

*I thank Susan Godlonton for excellent research assistance. Some parts of this paper were presented at seminars at SALDRU, Princeton and Michigan. I thank all participants for useful comments.

applications. Nevertheless there are several pitfalls to this "black box" approach. In this paper we highlight a number of difficulties. In particular we show that it is important to think carefully about whether it is plausible that the assets are, indeed, all proxying for the same thing. If the ownership of assets has independent effects then the asset index could provide misleading results.

We apply the asset indices to the subject of obesity. Obesity has been increasing across the world. In developed countries it has become one of the main public health issues. Nevertheless it has increased even in developing countries. Many South Africans, even poor ones, have a high body mass (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). Indeed, it has been claimed that excess BMI is the fifth most important risk factor for chronic disease in South Africa, as measured by DALYs (Bradshaw et al. 2007, Table 1, p.646). Understanding some of the correlates of high body mass, in particular incomes, would therefore be useful. Unfortunately, as Filmer and Pritchett noted, the Demographic and Health Surveys, the largest available data sets with anthropometric information, do not have adequate socio-economic information.

In this paper we will therefore analyse the relationship between obesity and socio-economic variables on three publicly available data sets which have anthropometric information. Two of these are relatively small surveys that also have good socio-economic information. The third is the South African Demographic and Health Survey which only has asset information. We will use the first two to compare the performance of asset proxies to measured expenditure (or income). In the process we will highlight both the potential of asset proxies, as well as many of the pitfalls associated with their use.

The structure of this paper is as follows. In the next section we will review different methods for constructing asset indices. We highlight some of the conceptual and practical difficulties in this process. We then discuss different purposes for which these indices may be used and review some of the literature that has tried to compare different indices. In Section 4 we describe the techniques by which we intend to assess the performance of the proxy variables. We describe our three data sets in Section 5 and report our regressions in Section 6. Section 7 concludes.

2 Constructing Asset Indices

2.1 The Sum of Assets

One of the earliest approaches to using the assets information was simply to sum the number of assets that people had. This procedure does not have much to recommend it, other than ease of use. The idea that very different assets should all be weighted equally is not very attractive on theoretical grounds. It can also have paradoxical effects when certain assets are "inferior goods", so that their ownership makes households look more affluent when in reality it might signal less affluence. The other approaches discussed below all weight assets differently, letting the correlation structure between the assets determine which assets should count for more.

2.2 The Principal Components Approach

The idea of using the first principal component of a set of asset variables as an index for “wealth” or long-run income has been around in the social science literature for a while (McKenzie 2005, p.232), but its use has become more widespread after publication of the influential paper by Filmer and Pritchett (2001). The basic idea of principal components is to find the linear combination of the asset variables that maximisises the variance of this linear sum. What does this actually mean?

Figure 1 presents a simple bivariate case. The first principal component (indicated by the solid line) is that line which minimises the residual variance when this is measured **perpendicular** to the component. In this case the distance is measured parallel to the second principal component. If there are more than two variables involved, the first principal component will be the line that is oriented through the scatter in such a way that the residual variance in **any** direction perpendicular to the line will be minimised. As Figure 1 shows, the line defined by the first principal component can usefully be compared to the regression lines of one of the variables on the rest. These regression lines also minimise a sum of squared deviations, but they treat the variables asymmetrically – the deviations are measured along the axis defined by the particular dependent variable. One might therefore think of the first principal component as the line of “best fit” where all variables are treated symmetrically.

Geometrically we can envisage the process as picking a set of new axes so that the coordinates of each “point” (x, y) will be given by $(score_1, score_2)$ such that the “$score_1$” values will have as large a variance as possible, i.e. one wants to pick the first axis in the direction of maximum spread. We show the new coordinates for a few points graphically in Figure 2, where the scatter is the same as shown in Figure 1.

More formally, if we have k random variables $a_1, \ldots, a_k$, each standardised to be of mean zero and variance one, the objective is to rewrite these as

$$\begin{aligned}
a_1 &= v_{11}A_1 + v_{12}A_2 + \ldots + v_{1k}A_k \\
a_2 &= v_{21}A_1 + v_{22}A_2 + \ldots + v_{2k}A_k \\
&\vdots \\
a_k &= v_{k1}A_1 + v_{k2}A_2 + \ldots + v_{2k}A_k
\end{aligned} \tag{1}$$

where A_i are unobserved components, created so as to be orthogonal to each other. Writing this in vector notation as

$$\mathbf{a} = \mathbf{VA}$$

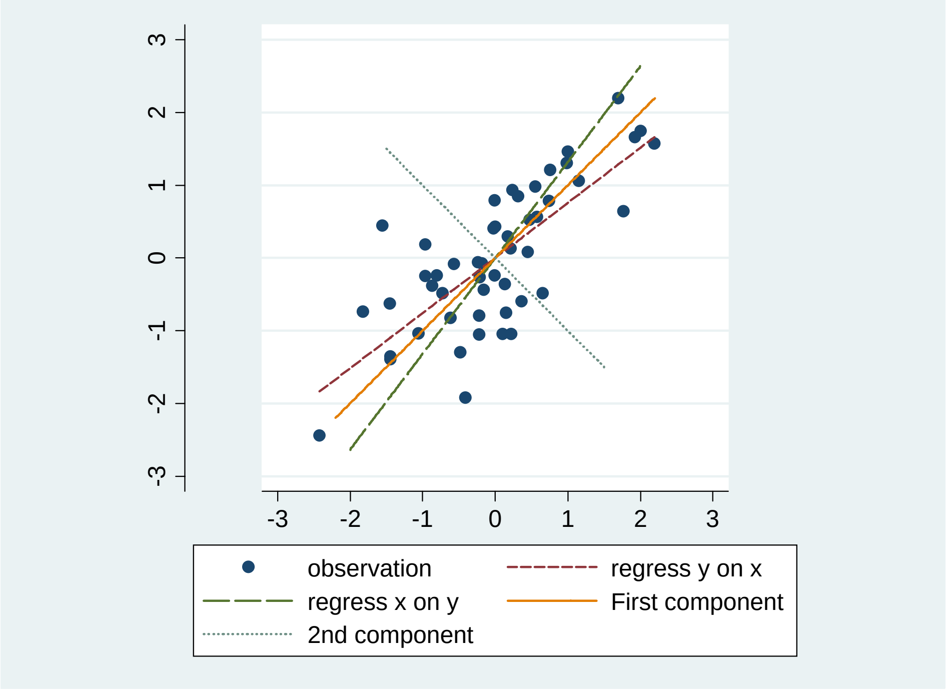

Figure 1: The regression line of y on x minimises the **vertical** sum of squared deviations; the regression line of x on y minimises the **horizontal** sum of squared deviations; while the line corresponding to the first principal component minimises the **perpendicular** sum of squared deviations.

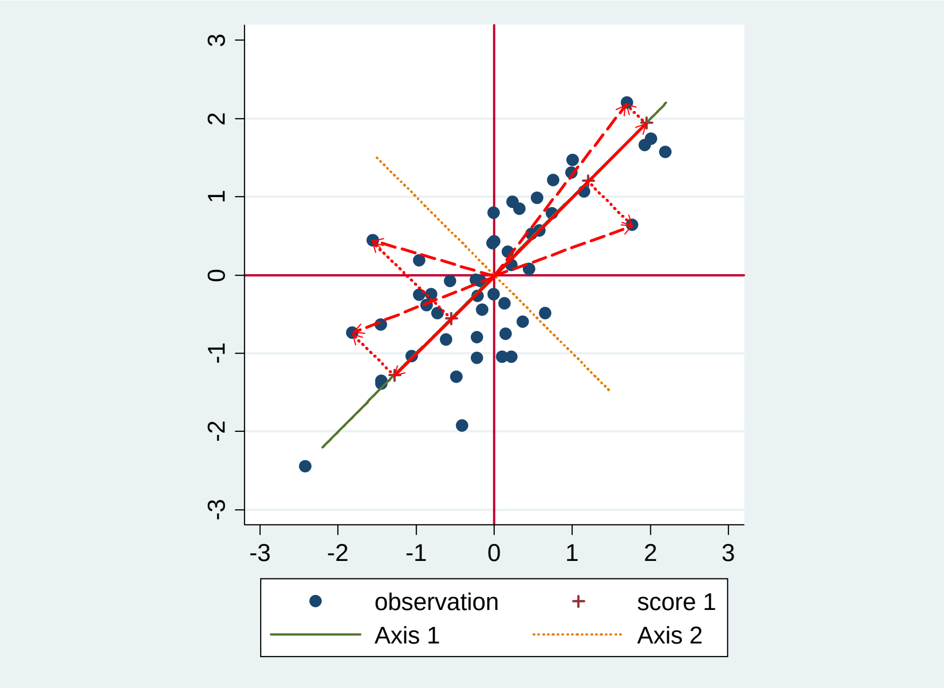

Figure 2: Each observation (x, y) can be rewritten as $(score_1, score_2)$ in terms of the new axes defined by the principal components.

it follows that the covariance matrix (here equal to the correlation matrix $\mathbf{R}$) is given by

$$\begin{aligned} E(\mathbf{aa}') &= E\left(\mathbf{VAA}'\mathbf{V}'\right) \\ \mathbf{R} &= \mathbf{V\Phi V}' \end{aligned}$$

where the last step follows from the fact that the $\mathbf{V}$ matrix is a matrix of fixed coefficients and we have let Φ be equal to $E\left(\mathbf{AA}'\right)$. Note that Φ is diagonal since the unobserved components are assumed to be orthogonal to each other. It is evident that we need to impose some normalisation in order to get a determinate solution. The obvious normalisation is to let Φ be the matrix of eigenvalues and $\mathbf{V}$ the orthonormal matrix of eigenvectors. Assume that $\mathbf{V}$ is ordered so that the eigenvector associated with the largest eigenvalue is listed first. We can then solve for $\mathbf{A}$ to get

$$\mathbf{A} = \mathbf{V}'\mathbf{a}$$

in particular

$$A_1 = v_{11}a_1 + v_{21}a_2 + \ldots + v_{k1}a_k \tag{2}$$

Note that $v_{11}^2 + v_{21}^2 + \ldots + v_{k1}^2 = 1$ since these are just the elements of the first eigenvector. Furthermore we have $var\left(A_1\right) = \lambda_1$, the first and largest of the eigenvalues since Φ is the diagonal matrix of eigenvalues.

Since $tr(\mathbf{R}) = k = tr(\Phi)$, a measure of the total variance explained by A_1 is given by λ_1/k. It is fairly easy to show (see appendix) that $var(\mathbf{a}'\mathbf{b}) \leq \lambda_1$ for **any** other set of coefficients $\mathbf{b}$ standardised so that $\|\mathbf{b}\| = 1$. It is in this sense that the first principal component gives that linear combination of the variables that maximises their joint variance.

If the asset variables a_i do not have unit variance and zero mean, they are first standardised, so that the equation for the first principal component will be given by

$$\begin{aligned} A_1 &= v_{11}\left(\frac{a_1 - \overline{a}_1}{s_1}\right) + v_{21}\left(\frac{a_2 - \overline{a}_2}{s_2}\right) + \ldots + v_{k1}\left(\frac{a_k - \overline{a}_k}{s_k}\right) \\ &= \frac{v_{11}}{s_1}a_1 + \frac{v_{21}}{s_2}a_2 + \ldots + \frac{v_{k1}}{s_k}a_k - c \end{aligned} \tag{3}$$

where the coefficients v_{i1} are the elements of the eigenvector $\mathbf{v}_1$ associated with the largest eigenvalue λ_1 of the correlation matrix $\mathbf{R}$ of the a_i variables. The constant c is the weighted sum of the means, which ensures that A_1 has a zero mean.

A consideration of equation 3 is useful, because it throws light on the claim that

> PCA works best when asset variables are correlated, but also when the distribution of variables varies across cases, or in this instance, households. It is the assets that are more unequally distributed between households that are given more weight in PCA (McKenzie 2003). Variables with low standard deviations would carry a low weight from the PCA; for example, an asset which all households own or which no households own (i.e. zero standard deviation) would exhibit no variation between households and would be zero weighted, and so of little use in differentiating SES. (Vyas and Kumaranayake 2006, p.461)

In the case of continuous variables this statement is plainly mistaken. The "loading" coefficients v_{i1} do not depend on the variances (since these have all been standardised to one in $\mathbf{R}$) but only on the correlations and since s_i is in the denominator, variables with low standard deviations will actually receive a higher weight. Variables with zero standard deviation cannot be standardised and so will not feature in the index. In the case of binary variables, however, we note that the variance is **largest** when the underlying probability is close to zero or one. It is therefore true that assets that are highly unequally distributed will be weighted more heavily in the index, but only if they are captured by means of a binary variable.

The fact that principal components are based largely on the structure of the correlations can be a limitation. Consider the simple "structural" model

$$a_1 = z + w + \nu_1 \tag{4a}$$

$$a_2 = z + w + \nu_2 \tag{4b}$$

$$a_3 = z + \nu_3 \tag{4c}$$

where z is unobserved permanent income, w is another common factor (e.g. "urbanisation") and the ν terms are idiosyncratic errors. Assume that z, w and ν are all uncorrelated and the variances are $var(z) = 0.4$,

$var\left(w\right) = 0.4$, $var\left(\nu_1\right) = var\left(\nu_2\right) = 0.2$, $var\left(\nu_3\right) = 0.6$. The eigenvector corresponding to the largest eigenvalue is $\left[\begin{array}{ccc} .627\,96 & .627\,96 & .459\,7 \end{array}\right]'$, i.e. the first two assets get weighted 36% more in the first principal component than the last asset. It is easy to show, however, that the linear index that maximises the correlation with the true z variable is one which weights all assets equally. In this case principal components picks up not only the common permanent income, but also a part of the common "urbanisation" variable.

2.3 The Factor Analysis Approach

The attractiveness **and** limitation of principal components analysis is that it can be seen as a purely descriptive device. A related approach which is more structural in intent is factor analysis. In this case the equation system (1) is modified as

$$\begin{aligned} a_1 &= v_{11}A_1 + v_{12}A_2 + \ldots + v_{1q}A_q + \nu_1 \\ a_2 &= v_{21}A_1 + v_{22}A_2 + \ldots + v_{2q}A_q + \nu_2 \\ &\vdots \\ a_k &= v_{k1}A_1 + v_{k2}A_2 + \ldots + v_{kq}A_q + \nu_n \end{aligned} \tag{5}$$

with $q < k$. It is assumed that the error terms ν_i are mutually orthogonal. We can write this system in matrix form as

$$\mathbf{a} = \mathbf{VA} + \boldsymbol{\nu}$$

The covariance matrix Σ of the asset variables is now given by

$$\Sigma = \mathbf{V}\Phi\mathbf{V}' + \Psi$$

where Ψ is the diagonal error covariance matrix and Φ is the covariance matrix of the underlying factors A_i. Note that in this case $\mathbf{V}$ is a $k \times q$ matrix, i.e. it is not of full rank. Furthermore there is no presumption that Φ need be diagonal. In order to estimate this model quite a lot of additional structure needs to be imposed. Indeed since everything on the right hand side of equation 5 is unobserved, there are infinitely many solutions as it stands. In order to get a definite solution we can rescale the $\mathbf{A}$ variables so that they all have unit variance, e.g. Let $\mathbf{A}^* = \Phi^{-\frac{1}{2}}\mathbf{A}$, and $\mathbf{V}^* = \mathbf{V}\Phi^{\frac{1}{2}}$, then $E\left(\mathbf{A}^*\mathbf{A}^{*'}\right) = \mathbf{I}_q$, while $\mathbf{a} = \mathbf{V}^*\mathbf{A}^* + \nu$. So we can make the assumption that $E\left(\mathbf{AA}'\right) = \mathbf{I}_q$ without undue loss of generality. In this case we can write the covariance matrix as

$$\Sigma = \mathbf{VV}' + \Psi \tag{6}$$

However even with this restriction there are still infinitely many solutions. Indeed given any particular solution for $\mathbf{A}$ and $\mathbf{V}$ we can generate valid new solutions $\mathbf{A}^* = \mathbf{HA}$ and $\mathbf{V}^* = \mathbf{VH}'$ for any set of orthogonal matrices $\mathbf{H}$ and $\mathbf{H}'$, i.e. matrices such that $\mathbf{HH}' = \mathbf{I}$. Such matrices represent "orthogonal rotations" and once an initial solution has been found, the analyst is free to "rotate" the solutions to get "interpretable" factors. In essence the analyst is free to choose the axes within the space spanned by the vectors $\mathbf{v}_i$. Indeed there are also "oblique rotations" (Kim and Mueller 1978, pp.37ff) which relax the requirement that the factors be orthogonal to each other. This fundamental indeterminacy has reduced the attractiveness of factor analysis to economists.

In order to get an initial definite solution orthogonality is imposed, plus the additional requirement that $\mathbf{V}'\Psi^{-1}\mathbf{V}$ be diagonal (Krzanowski 2000, p.483). One approach to producing estimates of the loading matrix $\mathbf{V}$ is an iterative one (Krzanowski 2000, p.487–488): begin with an initial estimate of Ψ and create $\mathbf{S} - \widehat{\Psi}$, where $\mathbf{S}$ is the sample covariance matrix. Then extract the first q eigenvectors to generate an initial estimate of $\mathbf{V}$. This then generates a new set of estimates for Ψ as the diagonal of $\mathbf{S}-\widehat{\mathbf{V}}\widehat{\mathbf{V}}'$. This then leads to a new matrix $\mathbf{S}-\widehat{\Psi}$ and so on. The process however, may not converge very rapidly or at all. In practice maximum likelihood estimation is much better (Krzanowski 2000, p.488). This, however, requires us to assume multivariate normality of the latent variables A_i and the errors ν. This is typically a bad assumption given the sort of assets that are often used, as we discuss in more detail below.

Once the factor coefficients $\mathbf{V}$ and the error variance matrix Ψ have been estimated, the task is to generate estimates of the factors themselves. This cannot be done by simple inversion of the formula as it was done in the case of PCA. Firstly the matrix $\mathbf{V}$ is not of full rank and secondly there is the unobservable error ν to contend with. Despite these differences, factor analysis still tries to approximate the latent variables by a linear combination of the asset proxies. This is done, for instance, by projecting $\mathbf{A}$ onto the assets $\mathbf{a}$, i.e. we write

$$A_i = \gamma_{i1}a_1 + \gamma_{i2}a_2 + \ldots + \gamma_{ik}a_k + \eta_i$$

where η_i is picked so as to be orthogonal to the asset proxies. This projection will exist, even though it may have no causal or structural interpretation. In vector form

$$\mathbf{A} = \Gamma\mathbf{a} + \boldsymbol{\eta}$$

Now

$$E\left(\mathbf{A}\mathbf{a}'\right) = \Gamma E\left(\mathbf{a}\mathbf{a}'\right) + E\left(\boldsymbol{\eta}\mathbf{a}'\right)$$

Since we are forcing $E\left(\boldsymbol{\eta}\mathbf{a}'\right) = \mathbf{0}$ and we have $E\left(\mathbf{A}\mathbf{a}'\right) = \mathbf{V}'$, and $E\left(\mathbf{a}\mathbf{a}'\right) = \mathbf{V}\mathbf{V}' + \Psi$, we have the following relationship involving the population projection coefficients Γ

$$\begin{aligned} \mathbf{V}' &= \Gamma\left(\mathbf{V}\mathbf{V}' + \Psi\right) \\ \Gamma &= \mathbf{V}'\left(\mathbf{V}\mathbf{V}' + \Psi\right)^{-1} \end{aligned}$$

so a method of moments estimator for the assets would be

$$\widehat{\mathbf{A}} = \widehat{\mathbf{V}}'\left(\widehat{\mathbf{V}}\widehat{\mathbf{V}}' + \widehat{\Psi}\right)^{-1}\mathbf{a} \tag{7}$$

where, of course, $\widehat{\mathbf{V}}\widehat{\mathbf{V}}' + \widehat{\Psi}$ could be estimated simply through the sample covariance matrix of the assets $\mathbf{S}$.

Sahn and Stifel (2003) have used this approach to estimate asset indices for a range of developing countries. They check how well these indices compare to expenditures. In a related piece (Sahn and Stifel 2000) they use asset indices created by factor analysis to look at changes in welfare over time, and to compare poverty between eleven countries in Africa. Their paper, however, highlights one of the key difficulties in getting factor analysis to work in this context: in order to avoid some of the indeterminacies outlined above, they posit that there is only one common factor in all of the variables. This implies that the variables will be uncorrelated after controlling for this "welfare" or "permanent income" variable. This, however, is hardly likely to be the case, considering that some of the variables used are infrastructure ones such as household sanitation and water supply. Furthermore ownership of electric appliances is also likely to be correlated with another common variable, i.e. access to the electric grid. If one includes multiple factors, however, the analysis becomes more messy. Firstly one has to define how many factors to include in the analysis and secondly one has to decide how to "rotate" the solution after extracting the initial factors. As a result the final "asset index" would be dependent on the particular choices made by the analyst as much as on the data.

In theory factor analysis might back out the latent variables in a system such as equations 4a-4c. In this case it turns out that it wouldn't: the correlation matrix can be reproduced perfectly with one latent variable $\boldsymbol{\xi} = \mathbf{z} + \mathbf{w}$; the variable $\mathbf{a}_3$ loads on $\boldsymbol{\xi}$ with factor $\frac{1}{2}$. The asset index (equation 7) would be proportional to $8\mathbf{a}_1 + 8\mathbf{a}_2 + \mathbf{a}_3$. The factor analysis would therefore create a hybrid "income plus urbanisation" variable even more strongly than principal components would. The reason for this is that factor analysis tries to ensure that the residuals are uncorrelated with each other.

Even in cases where distinct latent variables could be isolated more precisely, it would be uncertain in practice whether or not the analysis has succeeded. It would certainly need a careful argument to establish the validity of the "wealth" factor thus isolated.

2.4 Extracting an index from discrete data

There is an additional issue that is awkward for factor analysis. The assumption of multivariate normality required for the maximum likelihood estimation approach simply does not hold in the case of indicator variables; and most of the variables available for the construction of asset indices are of this type.

Several approaches have been suggested in the literature. Montgomery and Hewett (2005) adapt the equation system (5) in two ways: firstly, the equations are interpreted as latent variable equations for the corresponding indicator variables; and secondly the unique latent factor A_1 is itself modelled as

$$A_1 = \mathbf{W}'\boldsymbol{\theta} + u \tag{8}$$

where $\mathbf{W}$ contains variables that are thought to determine income (e.g. education) and u is a household specific idiosyncratic error. The latent variable equations can therefore be written (we have added in intercepts) as:

$$\begin{aligned}
a_1^* &= \alpha_1 + \mathbf{W}'\boldsymbol{\theta} + u + \nu_1 \\
a_2^* &= \alpha_2 + v_{21}\mathbf{W}'\boldsymbol{\theta} + v_{21}u + \nu_2 \\
&\vdots \\
a_k^* &= \alpha_k + v_{k1}\mathbf{W}'\boldsymbol{\theta} + v_{k1}u + \nu_k
\end{aligned} \tag{9}$$

Here the v_{11} coefficient has been normalised to be one. Assuming independence between $u, \nu_1, \ldots, \nu_k$, multivariate normality, and normalising each latent variable equation so that $var\left(v_{j1}u + \nu_j\right) = 1$, it then becomes possible to estimate the coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$ and $\mathbf{v}$ by maximum likelihood. Once these have been estimated one can estimate $\widehat{A}_1 = \widehat{E}\left(A_1|\mathbf{W},\mathbf{a}\right)$ by Gaussian quadrature (Montgomery and Hewett 2005, p.404). A related approach, using different normalisations and omitting the intercepts $\boldsymbol{\alpha}$ (but then estimating asset specific cut-points) is given by Ferguson, Tandon, Gakidou and Murray (2003).

In both cases we require the household random effect u to be uncorrelated with $\mathbf{W}$, as well as being uncorrelated with the asset specific errors $\nu_1, \ldots, \nu_k$ and the latter to be uncorrelated with each other. This assumption, in particular, is likely to be violated, for the same reason that we would expect different types of assets (household infrastructure, electrical devices) to show correlations even once the common effects of income or expenditure have been removed.

2.5 Proxy variable regressions

The Montgomery and Hewett (2005) approach makes an important distinction, that is too often ignored in the mechanical construction of asset indices: the difference between "causal" proxies, i.e. variables that belong on the right hand side of equation 8 and "outcome" proxies, i.e. variables that belong on the left hand side of the equations 9. Bollen, Glanville and Stecklov (2001) discuss the importance of this distinction in more detail. They point out that if the main purpose of the assets is to proxy for "income" in the estimation of a regression, then it matters what type of proxy it is.

To fix the issue, let us suppose that the main regression is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{z}\beta + \boldsymbol{\varepsilon} \tag{10}$$

where z is an unmeasured (or unmeasurable) variable, here thought of as "income". Because we have this particular variable in mind, we prefer this notation to the generic "factor" A_1. Assume now that we have a set of "causal proxies"

$$\mathbf{z} = \mathbf{W}\boldsymbol{\theta} + \mathbf{u}$$

with $\mathbf{u}$ uncorrelated with $\mathbf{W}$ and $\mathbf{X}$. In this case we can get unbiased estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}\beta$ by estimating the proxy variable regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\theta}\beta + \boldsymbol{\upsilon} \tag{11}$$

where $\boldsymbol{\upsilon} = \mathbf{u}\beta + \boldsymbol{\varepsilon}$ (for a discussion see Wooldridge 2002, pp.63–67).

If we have an outcome (or indicator) proxy, by contrast, i.e.

$$\mathbf{a}_1 = \rho_1 \mathbf{z} + \boldsymbol{\nu}_1$$

it follows that the model (10) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{a}_1 \frac{\beta}{\rho_1} - \frac{1}{\rho_1}\boldsymbol{\nu}_1 + \boldsymbol{\varepsilon} \tag{12}$$

and estimating this by OLS would lead to biased and inconsistent estimates due to the correlation between the proxy and the error term. Observe, however, that this does not mean that estimating regression 12 might be a pointless exercise. The bias might be smaller than if the variable had been omitted altogether, for instance. Furthermore if ρ_1 is known (or can be consistently estimated), then by the usual attenuation bias result, the OLS regression will give a lower bound on $|\beta|$. This might provide very useful information.

How do things change if we have more than one indicator proxy? Let us assume that we have k proxies

$$\begin{aligned} \mathbf{a}_1 &= \rho_1 \mathbf{z} + \boldsymbol{\nu}_1 \\ \mathbf{a}_2 &= \rho_2 \mathbf{z} + \boldsymbol{\nu}_2 \\ &\vdots \\ \mathbf{a}_k &= \rho_k \mathbf{z} + \boldsymbol{\nu}_k \end{aligned} \tag{13}$$

If the error terms $\boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_k$ are uncorrelated with $\boldsymbol{\nu}_1$ and with $\boldsymbol{\varepsilon}$, then regression (12) can be consistently estimated, using $\mathbf{a}_2, \ldots, \mathbf{a}_k$ as instruments for $\mathbf{a}_1$. However, if they are correlated then this strategy will also fail.

Some authors suggest one should simply put all the proxies into a regression of the sort

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{a}'\mathbf{b} + \boldsymbol{\zeta} \tag{14}$$

in the hope that the proxies will absorb the effect of the unmeasured income, and thus give better estimates of $\boldsymbol{\gamma}$. Furthermore Montgomery et al. (2000) show that a joint test of the hypothesis that $\mathbf{b} = \mathbf{0}$ can be used to test whether or not the true coefficient of $\mathbf{z}$ in regression 10 is zero. Nevertheless the multiple proxy regression 14 has also attracted criticism. Indeed Bollen et al. (2001, p.174) argue that estimating this regression would be a "bad choice" since it doesn't deal with the measurement error issue; and the common element $\mathbf{z}$ could create collinearity problems. Furthermore it is not clear how the influence of the underlying income variable could be reconstructed from the coefficients of the assets variables (Filmer and Pritchett 2001, p.116).

Lubotsky and Wittenberg (2006) show that this scepticism is misplaced. Indeed they show that an attenuated estimate of β can be obtained from regression 14 as

$$\widehat{\beta}_{LW} = \widehat{\boldsymbol{\rho}}'\widehat{\mathbf{b}} \tag{15}$$

provided that $\rho_1 = 1$. The $\boldsymbol{\rho}$ vector can be consistently estimated by a method of moments estimator

$$\widehat{\rho}_i = \frac{cov\left(y, x_i\right)}{cov\left(y, x_1\right)} \tag{16}$$

Furthermore they show that if any linear combination of the asset variables $\mathbf{a}_{index} = \mathbf{a}'\boldsymbol{\delta}$ (e.g. the PCA index given in equation 3 or the Factor Analysis one in equation 7) is used as a proxy for $\mathbf{z}$ in regression 10, i.e. if we estimate the regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{a}_{index}\beta + \xi$$

then this index will lead to a more attenuated estimate $\widehat{\beta}_{index}$, provided that $\mathbf{a}_{index}$ is rescaled so that the sign of the bias can be determined.

The requirement that $\rho_1 = 1$ is not as restrictive as it might look. In the examples below we project $\mathbf{a}_1$ on $\mathbf{z}$ in a data set where we do observe both and then use the estimate $\widehat{\rho}_1$ obtained from this auxiliary data set to rescale $\mathbf{a}_1$ to $\frac{1}{\widehat{\rho}_1}\mathbf{a}_1$. This procedure will give consistent estimates, provided that the two data sets are sampled from the same population; the variables $\mathbf{a}_1$ are measured equivalently in both survey instruments; and the measured $\mathbf{z}$ in the one data set corresponds to the latent variable $\mathbf{z}$ in the other. If $\mathbf{z}$ is unmeasurable in principle, then the normalisation $\rho_1 = 1$ implicitly fixes its scale.

2.6 The Regression Weighted Index

The Lubotsky-Wittenberg (LW) estimator reconstructs an estimate of β that is least subject to attenuation bias, if the errors $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \dots, \boldsymbol{\nu}_k$ in equation 13 are correlated with each other, but uncorrelated with $\boldsymbol{\varepsilon}$. Their procedure also implicitly constructs an asset index. It can be written as (Lubotsky and Wittenberg 2006, p.554)

$$\mathbf{a}_{LW} = \frac{1}{\widehat{\beta}_{LW}} \sum_{j=1}^{k} \mathbf{a}_j \widehat{b}_j \tag{17}$$

This index is not an all-purpose income or welfare proxy. Rather it captures that part of the common correlation between the asset variables that best explains the outcome in question.

3 What are asset proxies good for in practice?

The literature that has used asset proxies is already vast and growing. It has found application in the study of childhood cognitive development (Paxson and Schady 2007), educational attainment (Filmer and Pritchett 1999), the situation of orphans (Ainsworth and Filmer 2006), the impact of economic development on child health (Boyle et al. 2006), inequality in health care among the poor (Schellenberg et al. 2003), poverty trends over time and between countries in Africa (Sahn and Stifel 2000), among others. Within the existing literature there are certain types of uses which recur, e.g. asset proxies are used to generate wealth rankings and poverty measures; inequality measures; they are used as controls in regressions; or they are used to estimate the impact of “wealth” on the outcome of interest. In the process some of the benefits as well as the limitations of asset proxies have been discovered.

3.1 Generating wealth rankings

Asset indices are frequently used to generate wealth rankings within a country. Indeed the original Filmer and Pritchett (2001) article generated wealth quintiles from the asset index and this type of use has persisted, particularly within the health literature. A number of authors have tried to validate these rankings against external or internal benchmarks and the results have been somewhat mixed.

3.1.1 Correlation with consumption rankings

Filmer and Pritchett (2001, p.120) validated their household rankings against rankings based on consumption (normalised for household size as C/N^{α} where the economies of scale parameter α was set to 0.6). They found

> The Spearman rank correlations across households are .64 for Nepal ($p < .001$, $N = 3{,}372$), .56 for Indonesia ($p < .001$, $N = 16{,}242$), and .43 for Pakistan ($p <.001$, $N = 1{,}192$). Clearly, the degree of agreement among the different rankings varies across the countries. Generally, the smaller the α, the better the fit between assets and expenditure classifications; thus the asset index classification fits total household expenditures better than is reported and fits per capita expenditures worse than is reported.

A study using Malawi household data that compared different types of asset indices (including a PCA index) to household per capita expenditure found

> All of the indices have similar levels of misclassification between quintiles of the wealth index and quintiles of per capita consumption expenditure, with only approximately 30% of households in the same quintile and Kappa statistics of roughly 0.1. (Howe et al. 2008)

These results may not be as negative for asset indices as they appear, since it may be that assets pick up longer-run welfare better than consumption does.

3.1.2 Sensitivity to assets included

Houweling et al. (2003) tested the PCA index rankings for sensitivity to the assets included. They were concerned about the fact that some of the assets (particularly water and sanitation) might have independent effects on the outcome of interest, in particular child mortality. Consequently they created increasingly more narrow versions of the index. The base level (the “World Bank” index) included all the assets typically used in the PCA index, i.e. consumer durables, household infrastructure variables (water, sanitation), household quality measures and access to electricity. The first variant ("Index 1") excluded only water and sanitation, the second variant excluded also housing quality measures, while the third also excluded electricity, thus leaving only the consumer durables. In the case of Indonesia and Uganda 27% of households moved to a different quintile when using "Index 1" when compared to the base. Yet larger changes occurred when Index 2 and Index 3 were used.

The choice of assets to include in the index is therefore an important matter, but it may be constrained by the assets available in the survey and the question at hand. The important point made by Houweling et al. (2003) is that one needs to be cautious in contexts where assets may be doing double duty: proxying for income on the one hand, but independently affecting the outcome on the other.

3.1.3 Proportion of asset variance explained

A number of studies have commented on the low proportion of the common variance typically explained by the PCA index . For instance Houweling et al. (2003) note that for the ten countries that they studied,

> The proportion of variance between households in the ownership of assets that is explained by the WB index is quite low (between 12 and 20%). The percentage of explained variance increased upon exclusion of items from the index to an average of 35% in the third, shortest, alternative index including only consumer goods.

Howe et al. (2008) also note that the first component from the PCA typicallly explains less than 20% of joint variance of assets.

These criticisms, however, seem misplaced. Consider, for instance, a set of k asset variables where each asset can be written as

$$\mathbf{a}_i = \rho \mathbf{z} + \boldsymbol{\nu}_i \tag{18}$$

where the error terms $\boldsymbol{\nu}_i$ are mean zero, mutually independent random variables, independent of $\mathbf{z}$. Assume that $var\left(\mathbf{z}\right) = 1$ and $var\left(\boldsymbol{\nu}_i\right) = 1 - \rho^2$, so that $var\left(\mathbf{a}_i\right) = 1$. In this case the first principal component will weight them equally with weights $v_{i1} = \frac{1}{\sqrt{k}}$. The proportion of the total variance explained by the first principal component will be (see appendix A.2)

$$\frac{1}{k} + \frac{(k-1)}{k}\rho^2 \tag{19}$$

This **decreases** with k, reaching ρ^2 in the limit. However, the correlation between the asset index $\mathbf{a}_{index} = \sum_{j=1}^{k} \frac{1}{\sqrt{k}} \mathbf{a}_j$ and $\mathbf{z}$ converges to one as $k \to \infty$. This example shows that the proportion of total variance explained is not a useful metric in which to think about the performance of the asset index. It is the correlation between the index and the latent variable $\mathbf{z}$ that is of interest.

3.1.4 Discrete nature of asset indices

In most of the cases where asset indices are constructed, the underlying asset variables are discrete and typically binary. This means that any index constructed from them will take on only a finite number of values. The fewer the number of assets included in the index, the more pronounced this characteristic will be. For instance in the study of health risk factors by Blakely, Hales, Kieft, Wilson and Woodward (2005) the asset index took on 96 distinct values. It probably would have taken on considerably fewer if "educational status" had not been among the explanatory variables. As McKenzie (2005) shows, the PCA index is much better behaved if it includes at least some continuous variables (e.g. number of rooms).

3.1.5 Clumping in the distribution

If there are too few levels of the index, it is possible that the index does not allow one to create fine rankings between households. For instance Houweling et al. (2003) comment

> An extreme example is Chad, where, when using Index 2 and 3 it became impossible to distinguish between the poorest 59% of the population. The reason is that none of the households in this group owned durable consumer goods or electricity, the only items included in Index 2 and 3.

In middle-income countries, the asset indices may find it difficult to differentiate between the "middle class" and the "rich" since the sort of assets listed in these surveys (such as television sets, cars, refrigerators) are likely to be owned by both.

3.2 Generating poverty measures

Some authors have used the "wealth rankings" to generate poverty measures or poverty comparisons. There is an obvious standardisation problem here, since there is no obvious "poverty line" for an asset index. Some authors have been content to generate relative deprivation measures, for instance identifying the bottom 40% on the asset index as "poor", the next 40% as "middle" and the top 20% as "rich" (Filmer and Pritchett 1999, p.89). Other authors create asset indices by pooling over time or over regions and then compare how asset holdings have changed over time, or how they differ between regions (Sahn and Stifel 2000). This procedure can be misleading, however. We noted above that the first principal component will pick up not just the common "wealth" factor, but also anything else that most of the variables tend to have in common. The common factor isolated by factor analysis will do likewise.

In cases where there are subgroups (regions) which have distinctive patterns of asset holdings, this procedure is likely to bias the poverty measures. In the South African case (discussed in more detail below) we will see that ownership of sheep and cattle receives a **negative** weight in the first principal component. The reason for this is that ownership of most consumer durables is also strongly correlated with urban residence. Ownership of sheep and cattle is negatively correlated with these other assets. Consequently it receives a negative weight. Poverty comparisons **within** the rural area will therefore be skewed – individuals with livestock and no consumer durables will look **poorer** than individuals who own absolutely nothing. Obviously it will also skew poverty comparisons between rural and urban areas. Rural people will look poorer, simply because some of their assets are never held by urban people and that will lead to low or negative correlations with the consumer durables typically held by urbanites.

3.3 Inequality measures

McKenzie (2005) discusses how one can use an asset index to generate measures of inequality. The raw asset indices cannot be used for many of these measures, since many require non-negative numbers (e.g Theil index), whereas asset indices are normed to be mean zero. He suggests that one can get useful estimates of inequality within sub-groups (e.g. regions) using as measure

$$\frac{\sigma_i}{\sqrt{\lambda_1}} \tag{20}$$

where σ_i is the standard deviation of the asset index within the i-th group and λ_1 is the first eigenvalue, i.e. it measures the overall variance of the PCA asset index. The ratio in equation 20 therefore compares the variability within the sub-group to the global variability in the asset index. Its relative magnitude therefore provides information on whether that group has excess variability when compared to the entire population, or not.

As in the case of the wealth rankings, this measure will only work reasonably if the distribution of the asset scores is not too concentrated. In cases like that reported for Chad above, where 59% of households obtained the same score, σ_i is likely to understate the true variability. The inequality measure may therefore end up assigning higher inequality to groups where the assets are better able to separate out different levels of well-being.

McKenzie also suggests that one can get estimates of *consumption* inequality from the asset indices by imputing (log) consumption values on the basis of the assets. This requires that one have an auxiliary survey available on which one can calibrate the relationship between assets and consumption. One issue that has to be addressed in that context is how to deal with the regression error. McKenzie recommends a bootstrapping approach and suggests that estimates of inequality based on this procedure do well, particularly if the Atkinson A(2) measure is used. He suggests that this might be expected if the asset indicators are better at distinguishing among the poor than among the rich (McKenzie 2005, pp.249-50).

3.4 Test of the relevance of income in a regression

Montgomery et al. (2000) have argued that a joint test of all the proxies in a regression like (14) would give a valid test of the hypothesis that the coefficient on income is zero. In implementing this test, however, they point out that one needs to decide whether any of the asset variables have a directly causal role. For instance, in a child mortality regression it is highly likely that water and sanitation play a direct role instead of merely proxying for income. Such asset variables theoretically belong among the **X** variables in regression 14 and not in **a**.

The authors compare the coefficients of the **X** variables in the proxy variable regressions to the coefficients that are obtained if consumption per adult is used instead. They come to the conclusion that in the cases that they consider the proxies manage to control for income quite well, so that the coefficients on the other variables of interest (in particular maternal education) are estimated reasonably well.

3.5 Estimating the effect of income or wealth

In most cases, however, the authors are interested in the effect of income or wealth itself. Some times this is done by dividing the households into asset quintiles and then running the estimation procedures separately by quintiles. The differences are then interpreted as being due to income. As Houweling et al. (2003) note, this procedure can be somewhat misleading if there are asset variables in the index that have direct causal impacts. They show that if water and sanitation are removed from the asset index, the measured inequality in child mortality is reduced. They conclude:

> For explanatory studies, though, it can be important to analyse the different sets of asset items separately, and not to combine them into one index. It enables the assessment of the relative importance of different components of material wealth, especially water and sanitation versus housing versus consumer items versus indicators of community wealth. Estimates of the relative importance of these components can contribute to the detection of causal mechanism that are most responsible for high child mortality among poor families.

These conclusions apply also to cases where the regression involves the asset index itself, instead of asset quintiles. If the asset index is correlated with a variable that **ought** to be in the regression (e.g. access to water) but is excluded, then the coefficients will obviously be biased and inconsistent. One would therefore want to include both an asset index proxying for the latent variable income as well as any of the assets that have direct effects. This raises an additional issue: does one estimate the asset index including or excluding the assets having direct effects?

This issue turns out to be a bit complicated, so we leave the details to an appendix (A.3). The double counting involved in including the asset twice will obviously bias the coefficients. Nevertheless there are other biases due to measurement error which are affected by whether or not the asset is included in the index. Despite this caveat, it seems to be a bad idea to deliberately bias the coefficients in the hope that the other biases might be diminished or counteracted, particularly since we cannot sign these other biases more precisely.

4 Testing the asset proxies

The literature review suggests that asset proxies have performed on the whole reasonably well. Indeed some types of analyses would not be possible without such indices. Nevertheless the existing literature also points to some of the dangers in using these indices, particularly when the assets involved also have direct effects.

In our empirical work we will be concerned with comparing regressions in which a "Filmer-Pritchett" style asset index is used as an explanatory variable to regressions in which we directly use income or expenditure. We will be concerned with the question whether the assets distort the regression results, particularly in terms of the estimated impact of the covariates. We will also consider whether the proxies give us an adequate lower bound on the true impact of the omitted expenditure variable, as suggested by Lubotsky and Wittenberg. Finally we will test to see whether the assets do, indeed, proxy for this latent variable or whether they seem to have direct effects. The test that we will use is the specification test outlined by Wittenberg (2007).

The intuition behind this test is quite straightforward. Assume that the "structural" model is given by 10; that the structural variable z is correlated with the vector of explanatory variables $\mathbf{X}$; and that the proxy variables $\mathbf{a}_k$ can be written in the form given in equation 13. The correlation between $\mathbf{z}$ and the assets will induce a correlation between the assets and $\mathbf{X}$ and, indeed, between the assets and the outcome $\mathbf{y}$. If the "measurement error" part of the asset proxies is not correlated with the regression error or the covariates then the structure of the correlations will overidentify the correlation coefficients ρ_k. An overidentification test can check to see whether using different "instruments" will lead to the same estimates of the parameter ρ_k. Wittenberg (2007) shows that the test can be implemented proxy by proxy as well as for the system of proxies as a whole. He notes that the test can fail for many reasons, *inter alia*:

- A correlation between the regression error $\boldsymbol{\varepsilon}$ and any of the "measurement errors" $\boldsymbol{\nu}_k$. This implies that the proxy variable $\mathbf{a}_k$ is proxying not only for the latent variable $\mathbf{z}$, but should be in the main regression. This is the case of the "water quality" measures in the morbidity and mortality regressions considered earlier.

- A misspecification of any of the proxy variable equations, e.g. if the proxy variable is not correlated with the latent variable or is correlated with one of the other explanatory variables.

- A misspecification of the main regression, including the omission of other relevant variables correlated with the proxies.

However any of these would be reasons for being sceptical about the validity of regressions using asset indices.

5 Samples and Measures

In the empirical part of the paper we will explore the relationship between body mass and a range of control variables, including income and expenditure (where available); asset variables and asset indices (of the "Filmer-Pritchett" type); and other individual and household attributes. We will be doing our analyses on three different data sets. Two of them are relatively small surveys that have both asset information as well as reasonably good socio-economic data. However, they are limited by their sample size and limited geographic coverage. The third survey is the South African Demographic and Health Survey from 1998, which is a nationally representative survey with large sample size. Like all Demographic and Health Surveys, its socio-economic information is seriously deficient.

5.1 The Surveys

5.1.1 Langeberg Survey (SAIHS)

The "South African Integrated Household Study" was conducted by a team of researchers from *inter alia* the University of Cape Town, University of the Western Cape, Princeton University and Harvard. The survey was coordinated by the South African Labour and Development Research Unit (SALDRU) at the University of Cape Town and was conducted in 1999. The survey was run in three Magisterial Districts (Mossel Bay, Heidelberg and Riversdal) in the Western Cape province, which collectively make up the Langeberg health district. Altogether 294 households were interviewed with a range of instruments: detailed modules on household socio-economic information, individual information and health related information, including anthropometric data. A distinctive feature of the study design (which we will not exploit in these analyses) was that all adult members of the household were separately interviewed.

The Langeberg district is more urbanised than South Africa as a whole. Furthermore its demographic make-up is quite different. In terms of the old apartheid classifications, Black South Africans are in a minority (at 14%), with the bulk of the population classified as "Coloured" (SAIHS 2001). In order to get an adequate sample size, we pooled Coloureds and Black South Africans.

As shown in Table 1, we have 561 individuals in our overall estimation sample, i.e. adults aged 20 above with complete information on all of the variables used in the regressions. Of these 457 were Black and Coloured. The Black sub-sample was 176.

The summary statistics in Table 1 show that the average body mass index in this sample is in the overweight category. Indeed half of the sample is overweight (i.e.with a BMI in excess of 25) while around a quarter is obese (BMI of 30 or above). Black and Coloured women in particular are very heavy, with around a third being obese.

5.1.2 KIDS 1998

The KwaZulu-Natal Income Dynamics Survey (KIDS) was a re-survey of the Black and Indian sub-samples from the KwaZulu-Natal province of the 1993 Project for Statistics on Living Standards and Development (PSLSD), otherwise known as the 1993 SALDRU survey. The PSLSD was a national survey based on the model of the World Bank Living Standards Measurement Surveys. The survey instrument consisted of a variety of modules, such as Expenditures, Incomes, Labour Market status as well as anthropometrics. Since the 1993 survey was not designed as a panel, the re-survey had to confront the problem of how to track individuals and households. The decision was made to follow "core" household members only. These were the Head of Household or spouse or any adult members present in 1993. Altogether just over one thousand households were sampled on this basis. All individuals present in these households were interviewed. If "non-core" household members had left the household, no attempt was made to track them. The 1998 survey can therefore not be interpreted easily as being representative of the Black and Indian population of KwaZulu-Natal province.

There are some other issues around representativeness. The original 1993 survey only measured children of six years and younger. The 1998 re-survey also measured some of the adults. As shown in Table 2, however, the coverage of the adults was highly incomplete. Less than a third of the adults in the sample have usable anthropometric information. Nevertheless the total sample size with anthropometrics is larger than in the case of the Langeberg Survey. A comparison of the summary statistics for the sample as a whole and the estimation sample (in Table 2) suggests that the indviduals in the estimation sample are notably older (a mean age of 48 when compared to 39), have less education and are less likely to be employed (38% versus 50%).

From Table 1 it is also apparent that the individuals in the KIDS sample are noticeably heavier than the ones captured in the Langeberg Survey and the DHS. Indeed fully 70% of the sample is overweight, while around half of the Black women were obese.

5.1.3 DHS 1998

The Demographic and Health Survey follows the template of other DHSs, i.e. it has detailed modules on child bearing, contraception and attitudes to family planning. The instrument that we will be analysing is the Adult Health Questionnaire which has information on health seeking behaviour, clinical conditions, occupational health, health-related habits as well as anthropometrics. Its socio-economic information, by contrast, is rudimentary to say the least. In particular there is no information about incomes or expenditures. There is, however, information about assets in the household questionnaire. This allows us to construct an "asset index" (labelled asset1 in Table 1) as advocated by Filmer and Pritchett (2001). We will comment on the validity of this procedure in more depth below.

The other variable that is poorly measured is labour market status. The household roster contains one question whether the individual worked for pay in the last seven days. The adult health module has a question (in the occupational health section) asking "In the last 12 months, have you worked for payment?" There is no additional information that might enable one to determine whether an individual is unemployed or not economically active, or indeed whether an individual might be employed informally or seasonally. We have chosen to work with the looser (i.e. 12 month) definition of employment, to capture any casual or seasonal workers.

The chief strength of the DHS is its sample size. As shown in Table 1 the usable sample is an order of magnitude greater than in the case of the KIDS or the Langeberg survey. Consequently this sample provides a lot more power in investigating the relationships between the "economic" variables (such as the asset index) and body mass.

5.2 Comparing the data sets

The summary statistics in Table 1 provide an interesting view on the comparability of the data sets. It is striking that the Langeberg Survey and the DHS have similar mean BMI figures, obesity and overweight rates for the entire sample as well as the Black subsamples. Furthermore the age profiles and household sizes are also very similar. The Langeberg area shows lower levels of education, but higher levels of employment, which is consistent with the fact that it is a relatively urbanised population but in a non-metropolitan setting. The rural parts of the Langeberg district are commercial farms which have relatively high employment levels. The typical rural areas in the rest of South Africa are the ex-homeland areas which have quite low employment levels.

The KIDS data set is markedly different, both in the BMI values as well as in the household sizes. While the average age in the KIDS sample as a whole is comparable to the average ages recorded for the Langeberg and DHS estimation samples, the individuals with usable BMIs in the KIDS sample are around ten years older. Interestingly enough, the log expenditures values in the KIDS data set are comparable to those in the Langeberg survey. However while the Langeberg survey shows a considerable mismatch between the income and expenditure figures, this is not the case in the KIDS survey. Since we used the household aggregates calculated by the relevant research teams, instead of estimating these *de novo*, this may be due to differences in the imputation processes used in the two surveys.

6 Validating the use of asset proxies

6.1 The regressions using the Langeberg survey

The Langeberg and the KIDS data sets have expenditure, income and asset information. We can therefore explore how well the asset proxies perform. Table 3 provides an initial assessment for the Langeberg survey. In column one we have regressed BMI on log total household expenditure, including some controls for household composition and personal characteristics.

In order to interpret the coefficient on expenditure we need to remember that BMI is weight divided by height squared, so a one unit increase in log expenditure would increase weight by 4. 08 kg (9 lb) for a person of average height in South Africa, i.e. 1.62m (5 ft 3 inches). Furthermore the income distribution in South Africa as a whole (as measured in the 2000 Income and Expenditure survey) is such that a movement from the 25th to the 75th percentile is equivalent to 1.4 units on the log scale. If this regression coefficient were to apply to all South Africans, it would imply a difference in average weight of 5.7 kg (12.6 lb) between individuals of average height at those two percentiles. This is a significant increase, both statistically and materially.

A number of the other coefficients are also interesting. The large "female" coefficient should not come as a surprise, given the summary statistics shown in Table 1. At the average height this would suggest that women are 9.2kg (20lbs) heavier than the corresponding men. Black individuals are on average heavier than Whites and Coloureds. The quadratic in age suggests a peak at age 51.

In column 2 we provide the comparison with the log of household income. As Table 1 shows, there are a substantial number of individuals (22%) that live in households that reported zero income. In order not to lose these from the sample, we set these incomes at R1 and included a separate dummy variable for this category. The large coefficient on the dummy suggests that individuals that live in these households are better off than their zero income would suggest - at least when measured in terms of their girth!

Interestingly enough the coefficient on the log of income is almost 40% smaller than the coefficient on log expenditure. The most plausible explanation is that log income is a more noisy measure of the real resources available to the individual. Indeed it is often assumed that income is more poorly measured in household surveys in developing countries than expenditure (Deaton 1997). The summary statistics in Table 1 support that view. It appears that in most households expenditure is markedly higher than income, suggesting problems with recording all incomes.

In column 3 we have used the Filmer and Pritchett (2001) asset index, i.e. we have extracted the first principal component from a range of asset variables. Several points may be noted. Firstly the coefficient of 1.142 on the asset index is difficult to interpret. In Table 1 we observe that the standard deviation of the variable is 2.0, so a one standard deviation increase in the index would lead to a 2.284 unit increase in BMI, which is larger than the increase associated with a one standard deviation increase in log expenditure. Nevertheless the distribution underlying the asset index is not equivalent to that of the expenditure variable so. there is no reason to suppose that a standard deviation increase in the index is equivalent to a standard deviation increase in income or expenditure.

A second point is that the coefficients on the other covariates in the regression are remarkably similar. Use of the asset index does not qualitatively distort the conclusions that we would draw from the regressions. This is reassuring.

Finally we note that the regression with the asset index fits better (as measured by the R^2) than either the regression with expenditure or income. This suggests that there is important information in the asset variables that does not seem to be contained in these other measures of well-being.

Lubotsky and Wittenberg (2006) have argued that it is preferable to include all the proxies separately in the regression and aggregate the coefficients afterwards. In order to implement their procedure we need to standardise our estimates on one of the proxy variables. We have chosen telephone ownership for a number of reasons. Firstly, the bivariate correlation between telephone ownership and BMI is fairly strong. Secondly it is harder to imagine a direct impact of telephone ownership on BMI than would be the case with some of the other variables, such as television or car ownership. In order to make the coefficients directly comparable to the expenditure coefficients in column 1, we have rescaled the telephone variable, i.e. we projected telephone ownership on expenditure as

$$telephone = b_1 + b_2 \ln \text{ expenditure} + \varepsilon$$

and then rescaled the telephone variable as $\frac{1}{b_2} telephone$. The coefficient on this rescaled variable is given as the coefficient on "asset 1" in column 4 of Table 3. We see that it is statistically significant in the regression, which we interpret as confirmation that it is a suitable proxy. We reiterate that this coefficient should be directly comparable to the coefficients in column 1 or 2. Astonishingly this proxy performs reasonably when used in this way, by itself. The coefficient shows "only" 42% attenuation. Interestingly, this regression shows a better fit than either the expenditure or the income regressions.

When the other assets are added (in column 5), the rescaled telephone variable is still statistically significant, but we observe that a number of additional assets seem individually significant. This need not be evidence of the fact that these variables belong in the main regression, since they may only be capturing the impact of the omitted expenditure variable. The fit of the regression improves further still.

In column 6 we aggregate up the coefficients as suggested by LW[1]. The estimated coefficient of 1.617 is remarkably close to the "true" coefficient of 1.554. The fact that it is larger than this coefficient is troubling only if we assume that the assets are proxying for income/expenditure. If both are proxying for "permanent income" then it might not be surprising that the LW procedure gives a "better" estimate than either consumption or expenditure. LW also provide a procedure by which the coefficients on any other linear combination of the asset proxies can be compared with the LW estimator. The coefficient corresponding to the Filmer-Pritchett (FP) asset index is given in the last line of column 2. This coefficient of 1.328 is not a bad estimate of the expenditure effect, but compared to it the LW procedure manages to extract a stronger signal.

In summary the asset indices do a reasonably good job of proxying for expenditure or income in this particular example. The coefficients on the covariates are generally of the same sign and of similar magnitude. Furthermore with some manipulation, the coefficients on the asset indices yield effects that are similar to the measured expenditure impact.

[1] To be precise we use the more efficient version of the aggregation process described in Wittenberg (2007).

We might be content to leave things here. However, we show at the bottom of column 6 in Table 3 that the Wittenberg (2007) omnibus specification test soundly rejects the validity of the model. As we noted earlier, the test can be applied to individual asset proxies and these tests suggest that several variables such as television and car ownership do not function as simple proxies for the same omitted variable. The implication is that some of these variables have independent effects and belong in the main regression.

In column 7 we provide a regression that does pass the Wittenberg (2007) specification test. Note that these estimates are still based on the regression given in column 5. We have simply allowed some proxies to have independent effects, then reaggregated the coefficients of the remaining proxies. We have also allowed for some of the covariates to be correlated with the proxies, which requires us to correct their coefficients as well. The regression suggests that access to electricity has a sizable and significant impact on body mass. For an individual with the mean height (1.62m) the presence of electricity would increase weight on average by 4.6 kg (10lbs). The presence of television would add an additional 6kg (13lbs). The coefficient on car ownership, while not statistically significant, implies an increase in weight of 2.3kg (5.2lbs). The reason why we also display the coefficient for "bicycle" is that the specification test did not accept that bicycle ownership was proxying for income in the regression. In this case the reason is not that it has an independent effect (although in other contexts it might be expected to do so) but that it just does not seem to be a good marker of income.

The estimated coefficient for the latent variable is now half of its previous size. Nevertheless it still has a significant and meaningful impact.

In column 8 we show that these asset variables also appear significant even when we include the actual log expenditure variable, instead of recreating its impact through the other asset variables. The regression coefficients on the covariates are again very similar to those in the column with the proxies (column 7). Furthermore this revised "true" coefficient on log expenditure is only 15% higher, suggesting that the LW procedure does an excellent job in this particular case. The direction of the bias is also in line with the *a priori* expectations.

The final column estimates the same regression, but using the FP asset proxy. For this regression we did not recalculate the index stripping out the variables that were directly included, consequently those coefficients are subject to unknown biases as can be verified in the output. We would misattribute some of the impacts of television ownership and electricity to "income".

There are several lessons that flow from this particular example. Firstly we see that asset proxies, such as the FP index, can do a good job in capturing the impact of expenditure. Secondly we confirm the point noted in the literature review that asset proxies have to be used with care when some of the assets belong in the main regression. In this case it appears that certain assets which are associated with lifestyle changes, such as television, car ownership and access to electricity, have impacts on body mass that cannot simply be reduced to income effects. Nevertheless the remaining income effect can still be estimated reasonably well with the additional asset variables, even if we do not use a direct measure of income. Finally it seems important and worthwhile to test for whether the assets are, indeed, proxying for income.

6.2 The regressions using the KIDS survey

A contrasting picture is given by the KIDS survey, as shown in Table 4. In column 1 of that table, we observe that the coefficient on log expenditure is similar in the KIDS survey to the Langeberg survey (1.31 as against 1.55). Once again the impact of log income is markedly smaller, as shown in column 2. In this case the summary statistics do not suggest as big a problem as in the case of the Langeberg survey. Furthermore there was no problem with zero incomes. The much smaller coefficient implies that expenditure measures the flow of resources causing increases in weight much better than income, which may be more volatile (as shown, by the higher standard deviation).

The FP asset proxy (column 3) has a similar coefficient to that of log income although it cannot really be interpreted in the same way. As far as the coefficients on the covariates are concerned, the coefficients in column 1 give qualitatively similar results to those obtained with the asset proxy.

In the next three columns we repeat the exercise that we have already described in the case of the Langeberg survey. First (in column 4) we report a regression in which we have used only one proxy, again telephone access. This variable has been rescaled using log expenditure, so that the coefficient should be comparable with that in column 1. We notice that in this case the coefficient is only weakly significant and the attenuation seems massive. Once other proxies are added in (column 5), the telephone asset proxy ceases to be statistically significant or economically meaningful. This raises questions about whether it should be used to calibrate the expenditure effect at all.

The LW proxy, shown in column 6 still shows massive attrition. It only manages to capture one quarter of the true effect. The estimated coefficient implies a change in weight of 0.9kg (2lbs) for an adult of average height. As shown by the specification test at the foot of column 6, in this case too there is strong evidence that a number of the assets belong in the main regression.

Indeed in this case it proved impossible to find a specification that would pass the specification test! A consideration of the individual proxies suggested strongly that ownership of a refrigerator and of a television seemed to have strong independent effects. Car ownership and access to electricity also did not seem to proxy for "permanent income" in the same way as telephone ownership did.

Indeed part of the problem might have been that telephone ownership in this sample was simply not a good "anchor" proxy for expenditure. In fact the bivariate correlation between telephone ownership and log of household expenditure was only 0.2785 in this sample. Once key asset variables like ownership of a refrigerator, television set and a car are stripped out, there is not much "signal" left in the remaining variables.

The results shown in column 8 suggest, however, that it is correct that these assets should be stripped out. The coefficients on refrigerator and television ownership are meaningful and statistically signficant even if we use log expenditure as our measure of household resources. The coefficient on car ownership, although not statistically significant, is still sizable. It translates into a 1.62kg (3. 5lb) weight gain for a person of average height.

Once these assets are included in the main regression, the FP asset proxy does not add much (column 9). It is practically zero. Indeed the same is true of the LW estimate given in column 7.

It is clear that in this case the asset proxies are not as successful in capturing variation in affluence as they were in the Langeberg survey. Part of the reason may be that the sample is odd. We have noted that the KIDS sample does seem to be skewed when compared to the other data sets. Interestingly, however, the simple correlation coefficients between the FP asset index and log expenditure are almost identical in the two surveys: .59 in the case of Langeberg and .62 in the case of KIDS.

Before moving off the regressions in Table 4, it is useful to note that the poorer performance of the asset proxies did not lead to major distortions on the other coefficients in the models. Some of these coefficients are interesting. We observe that body mass increases in this sample with the number of children, the opposite of the effect in the Langeberg survey. The coefficient can be explained if children do many of the chores, allowing the adults to lead a more sedentary life. Otherwise one might have assumed that increases in household size would tend to reduce the resources available and so reduce body weight. The location coefficients are large, suggesting that urban living is associated with increases in body weight. It should be noted that the inclusion of these variables is not the reason why the asset proxies performed so poorly. Regressions without these variables led to only small increases in the coefficients of the asset proxies.

6.3 The regressions using the DHS

Unlike with the other two surveys, we cannot benchmark our estimates against regressions using income or expenditure. Nevertheless the evidence from the Langeberg survey and KIDS give us reasonable confidence that the asset indices will not give a distorted picture of the correlates of BMI. In Table 5 we provide a series of regressions with similar covariates to those used in the other two surveys. In this particular survey we also include a direct indicator of whether the individual was a smoker. This turns out to be highly significant. In this data set we could not calibrate our telephone proxy ("asset 1") against log expenditure directly. Instead we used the coefficient obtained when regressing telephone ownership on log expenditure in the Income and Expenditure Survey of 2000. As a result the coefficients in columns 2 through 5 of Table 5 can still be interpreted as lower bounds on the true expenditure effects.

The coefficients show some interesting similarities and differences with the other data sets. The "employed" in the DHS tend to be heavier, as indeed they are in the KIDS. In fact the coefficient is of a very similar magnitude. In the Langeberg survey the results are quite different. As Table 1 indicates, however, the Langeberg survey shows very high levels of employment compared to the rest of the country, which suggests that the characteristics of the employed and those not employed are likely to be somewhat different than elsewhere.

The quadratic in age shows qualitatively similar results in all three surveys. Body mass increases with age until a turning point is reached in the late forties or early fifties (Langeberg survey), or mid-fifties (KIDS and DHS).

Education shows the most variable pattern. In the Langeberg survey body weight decreases with education; in KIDS the coefficients are around zero, whereas they are positive in the DHS.

In all surveys there is a positive relationship with expenditure and assets. By far the strongest relationship is shown in the Langeberg Survey. It is not really meaningful to compare the principal components coefficients. The "raw" LW coefficients are strongest in the Langeberg Survey (1.617), weakest in KIDS (0.350) with the DHS coefficient intermediate between these two (0.650 in column 4 of Table 5). That coefficient translates into a difference of 2.4 kg (5. 3 lb) between the weights of average height individuals at the 25th and 75th percentiles of South Africa's income distribution.

In all surveys women are heavier than men, with the biggest difference shown in the KIDS data. Black South Africans (the base category in all regressions) are heavier than Whites, Coloureds or Indians.

The key point of Table 5, however, is that as with the other surveys the specification test reported at the base of column 4 strongly suggests that some of the assets belong in the main regression or that the regression is misspecified in some other way. In column 5 we have allowed a number of the proxies to have independent effects. As in the case of KIDS, television ownership and refrigerator ownership matter a lot. This is not altogether surprising. The former is implicated in lifestyle changes and reduction in more active forms of leisure, while the latter has an impact on the ready availability of food. In this case bicycle ownership is weakly statistically significant and it has the expected sign, i.e. tending to reduce weight while car ownership promotes weight gain. The positive relationship between sheep and cattle ownership on the one hand and weight gain on the other seems to be a straight-forward wealth relationship. Nevertheless it probably acts in a different manner to other wealth because it is a marker for rural and more traditional individuals.

However even after all these assets have been stripped out of the asset index, the coefficient on the index aggregated from the remaining proxies is still statistically significant, suggesting that income has a direct role on weight, not merely one mediated by the acquisition of labour-saving devices. Nevertheless the specification test still suggests that those proxies do not all act in the same way. Alternatively the regression may be misspecified in other ways. The regression in column 6 is provided merely for comparison. As noted before the coefficients on the proxies included twice will be biased. The coefficients that are comparable are those on the non-asset covariates. These are very similar.

7 Conclusions

Our review of the methods by which asset indices are calculated has highlighted the fact that one needs to be careful in thinking about what should be included. Any procedure which extracts a common signal (by means of principal components or factor analysis) could potentially also extract other common factors. This is likely to be a particular problem where certain categories of goods (e.g. electrical applicances) turn out to be good discriminators between the rich and the poor.

Thinking about the components of the asset index turns out to be particularly important if it is used as a control variable in a multiple regression. It is quite possible that some of the components need to be in the regression independently. In our empirical application it turned out that body mass was strongly related to ownership of labour-saving devices (car), food storage devices (refrigerators) and devices promoting a sedentary life-style (television). In the small data sets where we had access to expenditure measures some or all of these turned out to be important also.

This raises an interesting conceptual issue. Ownership of these assets is obviously not an exogenous fact. In a sense these are all income type of effects, but mediated through particular channels. Individuals who have a higher taste for labour saving devices at a given income will have a higher weight. People with similar tastes, but higher incomes, can indulge those tastes more. That is in essence what the regressions are picking up. To that extent it is important to separate these effects out. By lumping everything together in the asset index one may get a reasonable estimate of the total impact of wealth on the outcome, but one misses some of the nuances.

More positively, however, our results suggest that even where the regressions were misspecified (in the sense of leaving out asset variables that probably belonged there), the impact on the covariates was not catastrophic. Indeed the regressions with the asset indices provided reasonably robust pictures of the correlates of high body mass. In short asset indices can be highly useful tools for exploring socio-economic relations on data bases without good income information. They just need to be used with due care.

A Appendix: Derivations

A.1 Proof that $var\left(\mathbf{a}'\mathbf{b}\right) \leq \lambda_1$

We have $\mathbf{a} = \mathbf{V}\mathbf{A}$ so $\mathbf{a}'\mathbf{b} = \mathbf{A}'\mathbf{V}'\mathbf{b}$. Let $\mathbf{c} = \mathbf{V}'\mathbf{b}$ then $\mathbf{b} = \mathbf{V}\mathbf{c}$. We know that $\mathbf{b}'\mathbf{b} = 1$, i.e. $\mathbf{c}'\mathbf{V}'\mathbf{V}\mathbf{c} = \mathbf{c}'\mathbf{c} = 1$. Now $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = \mathbf{c}'\mathbf{A}$ and $var\left(\mathbf{a}'\mathbf{b}\right) = E\left(\mathbf{c}'\mathbf{A}\mathbf{A}'\mathbf{c}\right) = \mathbf{c}'\Phi\mathbf{c}$, i.e.

$$\begin{aligned} var\left(\mathbf{a}'\mathbf{b}\right) &= c_1^2\lambda_1 + c_2^2\lambda_2 + \ldots + c_k^2\lambda_k \\ &\leq c_1^2\lambda_1 + c_2^2\lambda_1 + \ldots + c_k^2\lambda_1 \\ &= \lambda_1 \end{aligned}$$

A.2 The proportion of variance explained vs correlation with the latent variable

Assume that we have k indicators, all of variance one where

$$\mathbf{a}_i = \rho\mathbf{z} + \boldsymbol{\nu}_i$$

where $var\left(\mathbf{z}\right) = 1$ and $var\left(\boldsymbol{\nu}_i\right) = 1 - \rho^2$, so that $var\left(\mathbf{a}_i\right) = 1$. The correlation matrix of these variables will be given by:

$$\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

Then the first principal component will weight them equally with weights $v_i = \frac{1}{\sqrt{k}}$. The variance of this linear combination will be

$$\begin{aligned} var\left(\sum \frac{1}{\sqrt{k}}\mathbf{a}_i\right) &= 1 + 2\sum_i \sum_{j<i} \frac{1}{k}\rho^2 \\ &= 1 + \frac{2}{k}\frac{(k-1)\,k}{2}\rho^2 \\ &= 1 + (k-1)\,\rho^2 \end{aligned}$$

(This is also the eigenvalue of the correlation matrix corresponding to any vector with equal elements.) Therefore the proportion of joint variance explained is

$$\frac{1}{k} + \frac{(k-1)}{k}\rho^2 \to \rho^2$$

Furthermore

$$cov\left(\frac{1}{\sqrt{k}}\mathbf{a}_i, \mathbf{z}\right) = \frac{k}{\sqrt{k}}\rho = \sqrt{k}\rho$$

and

$$corr\left(\frac{1}{\sqrt{k}}\mathbf{a}_i, \mathbf{z}\right) = \frac{\sqrt{k}\rho}{\sqrt{1 + (k-1)\,\rho^2}} \to 1$$

A.3 The impact of running a regression with an asset index and an individual asset variable

To be specific, assume that the structural equation (10) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{z}\beta + \mathbf{a}_1\phi + \boldsymbol{\varepsilon} \tag{21}$$

and we estimate this as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{a}_{index}\theta + \mathbf{a}_1\lambda + \boldsymbol{\xi} \tag{22}$$

where $\mathbf{a}_{index} = \mathbf{a}'\boldsymbol{\delta}$ is an index that contains $\mathbf{a}_1$. Consequently the last equation can be rewritten as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\gamma} + \left(\mathbf{a}_1\delta_1 + \mathbf{a}_2\delta_2 + \ldots + \mathbf{a}_k\delta_k\right)\theta + \mathbf{a}_1\lambda + \boldsymbol{\xi} \\ &= \mathbf{X}\boldsymbol{\gamma} + \mathbf{a}^*_{index}\theta + \mathbf{a}_1\left(\delta_1\theta + \lambda\right) + \boldsymbol{\xi} && \text{(23a)} \\ &= \mathbf{X}\boldsymbol{\gamma} + \mathbf{a}^*_{index}\theta + \mathbf{a}_1\phi + \boldsymbol{\xi} && \text{(23b)} \end{aligned}$$

where $\mathbf{a}^*_{index} = \mathbf{a}_2\delta_2 + \ldots + \mathbf{a}_k\delta_k$, i.e. $\mathbf{a}^*_{index}$ is an index without $\mathbf{a}_1$. Note that OLS estimation of regression 23b or 22 will produce identical results (provided that we note that $\phi = \delta_1\theta + \lambda$), since the former is simply a linear transformation of the latter. In order to see what happens when we estimate regression 23b we rewrite $\mathbf{a}^*_{index}$ in terms of the latent variable $\mathbf{z}$. Let us assume that the asset variables can be written as in equations 13. This means that

$$\begin{aligned} \mathbf{a}^*_{index} &= \mathbf{a}_2\delta_2 + \ldots + \mathbf{a}_k\delta_k \\ &= (\rho_2\mathbf{z} + \boldsymbol{\nu}_2)\,\delta_2 + \ldots + (\rho_k\mathbf{z} + \boldsymbol{\nu}_k)\,\delta_k \\ &= \mathbf{z}\,(\boldsymbol{\rho}^{*\prime}\boldsymbol{\delta}^*) + \boldsymbol{\nu}^{*\prime}\boldsymbol{\delta}^* \end{aligned}$$

where the starred quantities indicate that asset $\mathbf{a}_1$ has been excluded. Consequently the "structural" coefficient β in regression 21 will be equal to $(\boldsymbol{\rho}^{*\prime}\boldsymbol{\delta}^*)\,\theta$, i.e. $\theta = \beta\,(\boldsymbol{\rho}^{*\prime}\boldsymbol{\delta}^*)^{-1}$. Although this is the "structural" coefficient in the empirical regressions (23b and 22), its estimate will be attenuated due to the "measurement error" term $\boldsymbol{\nu}^{*\prime}\boldsymbol{\delta}^*$. This bias will be exaggerated by the fact that we know that $\mathbf{a}_1$ is also correlated with $\mathbf{z}$, so that the coefficient of $\mathbf{a}_1$ will pick up some of the effect of the missing $\mathbf{z}$ variable. This bias will typically be in the opposite direction to the attenuation bias, i.e. if $\beta\,(\boldsymbol{\rho}^{*\prime}\boldsymbol{\delta}^*)^{-1}$ is underestimated, then ϕ in the empirical regression 23b will tend to be overestimated. So the coefficient of $\mathbf{a}_1$ in regression 22 will be subject to **two** biases: the coefficient is $\phi - \delta_1\theta$, which will typically be an underestimate, but since ϕ will be overestimated, the overall direction of the bias is uncertain.

There is, however, an additional effect. If $\mathbf{a}_1$ is excluded from the calculation of the asset index *ex ante*, this is equivalent to setting $\delta_1 = 0$. However, the exclusion of $\mathbf{a}_1$ will also affect all the other weighting coefficients δ_i. There are two effects:

- the more assets are included in the estimation of the index, the better the index is – provided that the latent variable is the only factor common to the assets
- if there are other common factors (such as "urbanisation"), then inclusion of a particular asset could increase or decrease the bias in the index.

In short including $\mathbf{a}_1$ in the calculation of the asset index and then including both the asset index and $\mathbf{a}_1$ in the regression will have unpredictable results. It will affect the weights within the index itself (for good or ill), and then the coefficient on $\mathbf{a}_1$ will be subject to a double bias: one due to the measurement error in the asset index (which will itself be affected by whether or not $\mathbf{a}_1$ is included) and one due to the double counting of the variable itself.

References

Ainsworth, Martha and Deon Filmer, "Inequalities in Children's Schooling: AIDS, Orphanhood, Poverty, and Gender," *World Development*, 2006, *34* (6), 1099–1128.

Blakely, Tony, Simon Hales, Charlotte Kieft, Nick Wilson, and Alistair Woodward, "The global distribution of risk factors by poverty level," *Bulletin of the World Health Organization*, February 2005, *83* (2), 118–126.

Bollen, Kenneth A., Jennifer L. Glanville, and Guy Stecklov, "Socioeconomic Status and Class in Studies of Fertility and Health in Developing Countries," *Annual Review of Sociology*, 2001, *27*, 153–185.

___ , ___ , and ___ , "Economic Status Proxies in Studies of Fertility in Developing Countries: Does the Measure Matter?," *Population Studies*, 2002, *56* (1), 81–96.

Boyle, Michael H. et al., "The influence of economic development level, household wealth and maternal education on child health in the developing world," *Social Science and Medicine*, 2006, *63*, 2242–2254.

Bradshaw, Debbie et al., "Strengthening public health in South Africa: Building a stronger evidence base for improving the health of the nation," *South African Medical Journal*, 2007, *97* (8), 643–651.

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

Chou, Shin-Yi, Michael Grossman, and Henry Saffer, "An Economic Analysis of Adult Obesity: Results from the Behavioral Risk Factor Surveillance System," Working Paper 9247, NBER October 2002.

Deaton, Angus, *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore: Johns Hopkins University Press, 1997.

Ferguson, B.D., A. Tandon, E. Gakidou, and C.J.L. Murray, "Estimating Permanent Income using Indicator Variables," 2003. mimeo. World Health Organization, Geneva.

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.

___ and Lant Pritchett, "The Effect of Household Wealth on Educational Attainment: Evidence from 35 Countries," *Population and Development Review*, 1999, *25* (1), 85–120.

Houweling, Tanja A.J., Anton E. Kunst, and Johan P. Mackenbach, "Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter?," *International Journal for Equity in Health*, 2003, *2* (8).

Howe, Laura D., James R. Hargreaves, and Sharon R.A. Huttly, "Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries," *Emerging Themes in Epidemiology*, 2008, *5* (3). doi:10.1186/1742-7622-5-3.

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.

Kim, Jae-On and Charles W. Mueller, *Factor Analysis: Statistical Methods and Practical Issues* University Paper Series on Quantitative Applications in the Social Sciences, number 07-014, Newbury Park, CA: Sage, 1978.

Krzanowski, W.J., *Principles of Multivariate Analysis: A User's Perspective*, revised ed., Oxford: Oxford University Press, 2000.

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

McKenzie, David J., "Measuring inequality with asset indicators," *Journal of Population Economics*, 2005, *18*, 229–260.

Montgomery, Mark R. and Paul C. Hewett, "Urban Poverty and Health in Developing Countries: Household and Neighborhood Effects," *Demography*, 2005, *42* (3), 397–425.

___ , Michele Gragnolati, Kathleen A. Burke, and Edmundo Paredes, "Measuring Living Standards with Proxy Variables," *Demography*, 2000, *37* (2), 155–174.

Paxson, Christina and Norbert Schady, "Cognitive Development among Young Children in Ecuador: The Roles of Wealth, Health, and Parenting," *Journal of Human Resources*, 2007, *XLII*, 49–84.

Sahn, David E. and David Stifel, "Poverty Comparisons Over Time and Across Countries in Africa," *World Development*, 2000, *28* (12), 2123–2155.

___ and ___ , "Exploring Alternative Measures of Welfare in the Absence of Expenditure Data," *Review of Income and Wealth*, 2003, *49* (4), 463–489.

SAIHS, "Introduction to the SALDRU Langeberg 1999 Data Set," SALDRU, University of Cape Town 2001.

Schellenberg, Joanna Armstrong et al., "Inequities among the very poor: health care for children in rural southern Tanzania," *The Lancet*, February 15 2003, *361*, 561–66.

Vyas, Seema and Lilani Kumaranayake, "Constructing socio-economic status indices: how to use principal components analysis," *Health Policy and Planning*, 2006, *21* (6), 459–468.

Wittenberg, Martin, "Testing for a common latent variable in a linear regression," School of Economics, University of Cape Town. Available at http://mpra.ub.uni-muenchen.de/2550/01/MPRA_paper_2550.pdf 2007.

Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Mass.: MIT Press, 2002.

Table 1: Summary Statistics for the Estimation Samples

	Langeberg				DHS				KIDS			
Variable	All	Black & Coloured	Women	Men	All	Black	Women	Men	All	Black	Women	Men
bmi	26.7	26.5	28.7	24.1	25.9	25.9	27.7	23.3	28.8	29.2	30.6	26.2
	(7.57)	(7.83)	(7.11)	(7.89)	(6.59)	(6.54)	(6.85)	(5.10)	(6.52)	(6.58)	(6.67)	(5.30)
obese	0.24	0.22	0.33	0.09	0.22	0.22	0.32	0.09	0.36	0.39	0.50	0.15
	(0.42)	(0.41)	(0.47)	(0.29)	(0.42)	(0.42)	(0.47)	(0.28)	(0.48)	(0.49)	(0.50)	(0.36)
overweight	0.51	0.49	0.67	0.30	0.48	0.47	0.60	0.29	0.70	0.73	0.79	0.58
	(0.50)	(0.50)	(0.47)	(0.46)	(0.50)	(0.50)	(0.49)	(0.45)	(0.46)	(0.45)	(0.40)	(0.49)
Age	41.2	39.2	38.6	39.9	41.7	40.9	41.1	40.6	48.2	48.7	48.0	50.3
	(14.29)	(12.78)	(12.19)	(13.38)	(16.19)	(16.24)	(16.26)	(16.23)	(14.81)	(15.33)	(15.65)	(14.50)
household size	4.9	5.3	5.5	5.1	4.9	5.2	5.4	4.9	7.2	7.7	7.8	7.3
	(2.60)	(2.65)	(2.73)	(2.55)	(2.86)	(3.02)	(2.95)	(3.09)	(4.44)	(4.60)	(4.67)	(4.43)
years education	5.5	4.4	4.6	4.1	7.4	6.8	6.7	7.0	4.5	4.0	4.0	3.9
	(4.21)	(3.40)	(3.42)	(3.37)	(4.32)	(4.31)	(4.34)	(4.27)	(3.71)	(3.50)	(3.45)	(3.61)
ln HH Expenditure	7.3	7.1	7.1	7.1					7.4	7.2	7.2	7.2
	(0.87)	(0.71)	(0.69)	(0.72)					(0.81)	(0.68)	(0.68)	(0.69)
ln HH Income	5.7	6.0	6.0	5.9					7.4	7.3	7.2	7.4
	(3.20)	(2.89)	(2.79)	(3.00)					(1.06)	(1.02)	(1.03)	(0.98)
Zero income	0.22	0.17	0.16	0.19					0.0	0.0	0.0	0.0
	(0.41)	(0.38)	(0.36)	(0.39)								
FP asset index	0.1	-0.4	-0.3	-0.5	0.1	-0.6	-0.6	-0.5	0.2	-0.1	-0.2	0.0
	(2.00)	(1.76)	(1.73)	(1.78)	(2.00)	(1.56)	(1.58)	(1.53)	(1.65)	(1.51)	(1.50)	(1.52)
employed	0.61	0.64	0.55	0.74	0.39	0.34	0.25	0.45	0.38	0.35	0.27	0.52
	(0.49)	(0.48)	(0.50)	(0.44)	(0.49)	(0.47)	(0.44)	(0.50)	(0.48)	(0.48)	(0.44)	(0.50)
Pensioner in HH	0.14	0.14	0.14	0.14	0.30	0.31	0.33	0.29	0.36	0.40	0.43	0.34
	(0.40)	(0.38)	(0.38)	(0.39)	(0.46)	(0.46)	(0.47)	(0.45)	(0.48)	(0.49)	(0.50)	(0.47)
Number of adults	3.4	3.6	3.7	3.6	3.1	3.1	3.0	3.2	3.9	4.1	4.1	4.1
	(1.73)	(1.78)	(1.86)	(1.70)	(1.64)	(1.70)	(1.66)	(1.75)	(2.22)	(2.31)	(2.35)	(2.22)
Number of children	1.5	1.6	1.8	1.5	1.8	2.1	2.3	1.7	3.2	3.5	3.7	3.2
	(1.45)	(1.47)	(1.48)	(1.46)	(1.90)	(2.00)	(2.02)	(1.92)	(2.82)	(2.90)	(2.94)	(2.78)
smoker					0.30	0.26	0.08	0.50				
					(0.46)	(0.44)	(0.27)	(0.50)				
female	0.52	0.52			0.57	0.57			0.66	0.68		
	(0.50)	(0.50)			(0.49)	(0.49)			(0.47)	(0.46)		
black	0.31	0.39	0.38	0.39	0.73				0.85			
	(0.46)	(0.49)	(0.49)	(0.49)	(0.44)				(0.35)			
coloured	0.50				0.14							
	(0.50)				(0.35)							
asian/indian					0.04							
					(0.19)							
n	561	457	236	221	10299	7557	4342	3215	1444	1233	844	389

Table 2: Comparison between the estimation sample and the entire sample in KIDS

	All		Black		Women		Men		Estimation sample		Black		Women		Men	
Variable	Obs	Mean	Obs	Mean	Obs	Mean	Obs	Mean	Obs	Mean	Obs	Mean	Obs	Mean	Obs	Mean
bmi	1446	28.8	1235	29.2	845	30.6	390	26.3	1444	28.8	1233	29.2	844	30.6	389	26.2
		(6.52)		(6.58)		(6.67)		(5.30)		(6.52)		(6.58)		(6.67)		(5.30)
age	4266	38.9	3804	38.6	2196	39.5	1608	37.4	1444	48.2	1233	48.7	844	48.0	389	50.3
		(15.77)		(15.96)		(16.67)		(14.86)		(14.81)		(15.33)		(15.65)		(14.50)
household size	4682	7.8	4193	8.2	2357	8.3	1836	8.1	1444	7.2	1233	7.7	844	7.8	389	7.3
		(4.64)		(4.72)		(4.71)		(4.73)		(4.44)		(4.60)		(4.67)		(4.43)
years education	4633	5.5	4151	5.3	2355	5.2	1796	5.4	1444	4.5	1233	4.0	844	4.0	389	3.9
		(3.79)		(3.73)		(3.76)		(3.68)		(3.71)		(3.50)		(3.45)		(3.61)
FP asset index	4682	0.1	4193	-0.2	2357	-0.1	1836	-0.2	1444	0.2	1233	-0.1	844	-0.2	389	0.0
		(1.63)		(1.52)		(1.53)		(1.51)		(1.65)		(1.51)		(1.50)		(1.52)
ln HH Expenditure	4682	7.4	4193	7.2	2357	7.2	1836	7.2	1444	7.4	1233	7.2	844	7.2	389	7.2
		(0.75)		(0.66)		(0.66)		(0.65)		(0.81)		(0.68)		(0.68)		(0.69)
ln HH Income	4665	7.4	4178	7.3	2349	7.3	1829	7.3	1438	7.4	1229	7.3	841	7.2	388	7.4
		(1.05)		(1.02)		(1.02)		(1.02)		(1.06)		(1.02)		(1.03)		(0.98)
employed	4682	0.50	4193	0.50	2357	0.42	1836	0.61	1444	0.38	1233	0.35	844	0.27	389	0.52
		(0.50)		(0.50)		(0.49)		(0.49)		(0.48)		(0.48)		(0.44)		(0.50)
Pensioner in HH	4682	0.41	4193	0.44	2357	0.45	1836	0.43	1444	0.36	1233	0.40	844	0.43	389	0.34
		(0.49)		(0.50)		(0.50)		(0.49)		(0.48)		(0.49)		(0.50)		(0.47)
Number of adults	4682	4.3	4193	4.4	2357	4.4	1836	4.5	1444	3.9	1233	4.1	844	4.1	389	4.1
		(2.34)		(2.41)		(2.39)		(2.44)		(2.22)		(2.31)		(2.35)		(2.22)
Number of children	4682	3.5	4193	3.7	2357	3.8	1836	3.6	1444	3.2	1233	3.5	844	3.7	389	3.2
		(3.02)		(3.05)		(3.03)		(3.07)		(2.82)		(2.90)		(2.94)		(2.78)
female	4682	0.56	4193	0.56					1444	0.66	1233	0.68				
		(0.50)		(0.50)						(0.47)		(0.46)				
black	4682	0.90							1444	0.85						
		(0.31)								(0.35)						

Table 3: Assessing the performance of the asset proxies in the Langeberg survey

Dependent variable: BMI	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	ln Exp	ln Inc	PCA	Proxy Tel	Proxy All	LW	LW2	ln Exp2	PCA2
employed	-0.863	-1.014	-0.522	-0.478	-0.502	-0.502	-0.466	-0.738	-0.559
	[0.700]	[0.738]	[0.684]	[0.692]	[0.685]	[0.726]	[0.830]	[0.687]	[0.685]
age	0.326	0.351	0.347	0.37	0.368	0.368	0.368	0.327	0.351
	[0.124]**	[0.124]**	[0.121]**	[0.122]**	[0.122]**	[0.108]**	[0.106]**	[0.122]**	[0.122]**
age squared	-0.003	-0.003	-0.004	-0.004	-0.004	-0.004	-0.004	-0.003	-0.004
	[0.001]*	[0.001]*	[0.001]**	[0.001]**	[0.001]**	[0.001]**	[0.001]**	[0.001]*	[0.001]**
years education	-0.067	-0.011	-0.153	-0.075	-0.196	-0.196	-0.196	-0.181	-0.158
	[0.099]	[0.096]	[0.097]	[0.096]	[0.102]+	[0.080]*	[0.091]*	[0.102]+	[0.099]
Expenditure/ asset index/ asset 1	1.554	0.953	1.142	0.901	0.58	1.617	0.724	0.836	0.592
	[0.458]**	[0.346]**	[0.199]**	[0.201]**	[0.228]*	[0.296]**	[0.335]*	[0.480]+	[0.356]+
number adults	-0.057	0.053	-0.035	0.04	0.011	0.011	0.011	-0.105	-0.041
	[0.202]	[0.197]	[0.192]	[0.193]	[0.194]	[0.148]	[0.171]	[0.199]	[0.192]
number children	-0.459	-0.513	-0.325	-0.401	-0.396	-0.396	-0.396	-0.36	-0.343
	[0.231]*	[0.234]*	[0.228]	[0.229]+	[0.233]+	[0.235]+	[0.247]	[0.233]	[0.233]
female	3.562	3.59	3.4	3.398	3.436	3.436	3.436	3.469	3.425
	[0.625]**	[0.630]**	[0.613]**	[0.620]**	[0.613]**	[0.587]**	[0.659]**	[0.615]**	[0.615]**
white	-2.21	-1.579	-2.68	-2.732	-1.755	-2.732	-2.174	-2.295	-2.391
	[1.160]+	[1.123]	[1.119]*	[1.301]*	[1.104]	[1.210]*	[1.065]*	[1.238]+	[1.255]+
coloured	-2.697	-3.155	-2.5	-2.276	-2.226	-2.276	-2.189	-2.412	-2.485
	[0.714]**	[0.759]**	[0.700]**	[0.822]**	[0.711]**	[0.863]**	[0.888]*	[0.786]**	[0.790]**
zero income		5.768							
		[2.623]*							
electricity					2.129		1.756	2.25	1.108
					[1.077]*		[0.775]*	[0.926]*	[1.221]
television					2.417		2.312	2.436	1.966
					[0.788]**		[0.832]**	[0.778]**	[0.854]*
car					0.777		0.897	1.075	0.695
					[0.953]		[0.747]	[0.939]	[1.024]
bicycle					0.037		0.02	0.154	-0.043
					[0.718]		[0.632]	[0.703]	[0.724]
electric stove					-1.338				
					[0.980]				

coal stove					-0.831				
					[0.804]				
refrigerator					-0.124				
					[0.879]				
radio					1.106				
					[0.766]				
sewing machine					1.365				
					[0.921]				
motorcycle					-1.074				
					[2.294]				
Constant	9.726	13.008	20.951	16.643	18.353	16.643	16.643	11.849	18.505
	[3.762]**	[3.555]**	[2.866]**	[3.064]**	[2.851]**	[2.661]**	[2.675]**	[3.879]**	[3.189]**
Observations	561	561	561	561	561	561	561	561	561
R-squared	0.12	0.12	0.16	0.14	0.18	0.18	0.18	0.16	0.16
System test: Chi2						365.4	34.4		
df						90	30		
P value:						0	0.265		
adjusted coeff:			1.328						

Standard errors in parentheses

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

Standard errors in the LW regression have been bootstrapped, with 200 replications

Table 4: Assessing the performance of the asset proxies in the KIDS data set

Dependent variable: BMI	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	ln exp	ln inc	PCA	proxy tel	Proxy all	LW	LW2	ln exp2	PCA2
employed	0.279	0.089	0.208	0.32	0.244	0.244	0.306	0.269	0.264
	[0.376]	[0.382]	[0.378]	[0.367]	[0.379]	[0.354]	[0.376]	[0.376]	[0.378]
age	0.36	0.377	0.381	0.373	0.362	0.362	0.362	0.357	0.368
	[0.066]**	[0.066]**	[0.066]**	[0.059]**	[0.066]**	[0.060]**	[0.059]**	[0.066]**	[0.066]**
age squared	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003
	[0.001]**	[0.001]**	[0.001]**	[0.001]**	[0.001]**	[0.001]**	[0.001]**	[0.001]**	[0.001]**
years education	0.005	0.042	0.037	0.086	0.005	0.005	-0.003	-0.033	0.006
	[0.059]	[0.059]	[0.059]	[0.052]+	[0.060]	[0.054]	[0.052]	[0.060]	[0.059]
Expenditure/ asset index/ asset 1	1.309	0.720	0.778	0.123	0.094	0.350	0.103	0.810	0.070
	[0.280]**	[0.191]**	[0.217]**	[0.068]+	[0.063]	[0.107]**	[0.092]	[0.316]*	[0.457]
female	4.032	4.058	4.022	4.013	4.04	4.04	4.04	4.05	4.053
	[0.356]**	[0.357]**	[0.357]**	[0.315]**	[0.356]**	[0.312]**	[0.314]**	[0.355]**	[0.356]**
children	0.119	0.192	0.211	0.207	0.172	0.172	0.172	0.111	0.148
	[0.072]+	[0.070]**	[0.070]**	[0.080]**	[0.073]*	[0.079]*	[0.078]*	[0.073]	[0.072]*
number of adults	-0.042	-0.048	0.03	0.041	0.025	0.025	0.025	-0.021	0.029
	[0.088]	[0.090]	[0.086]	[0.090]	[0.086]	[0.088]	[0.089]	[0.088]	[0.086]
indian	-4.716	-3.77	-3.837	-3.189	-3.602	-3.602	-3.837	-4.717	-4.193
	[0.644]**	[0.601]**	[0.601]**	[0.656]**	[0.687]**	[0.754]**	[0.628]**	[0.669]**	[0.639]**
urban	1.107	1.035	0.733	1.082	0.655	0.655	0.705	0.913	0.95
	[0.513]*	[0.516]*	[0.546]	[0.568]+	[0.565]	[0.538]	[0.619]	[0.529]+	[0.551]+
city	1.791	1.865	1.67	1.989	1.575	1.575	1.741	1.749	1.938
	[0.626]**	[0.626]**	[0.642]**	[0.657]**	[0.659]*	[0.676]*	[0.686]*	[0.636]**	[0.647]**
electricity					-0.422		-0.259	-0.262	-0.318
					[0.443]		[0.414]	[0.415]	[0.561]
television					1.73		1.503	1.278	1.488
					[0.520]**		[0.582]**	[0.519]*	[0.546]**
refrigerator					1.361		1.217	0.972	1.238
					[0.441]**		[0.435]**	[0.444]*	[0.570]*
car					0.897		0.859	0.617	0.899
					[0.494]+		[0.433]*	[0.506]	[0.533]+
bicycle					0.36		0.226	0.092	0.226
					[0.523]		[0.533]	[0.518]	[0.521]

furniture					-0.406				
					[0.585]				
jewellery					-0.698				
					[0.349]*				
electrical appliance					0.001				
					[0.455]				
cattle					-0.347				
					[0.462]				
sheep					-0.053				
					[1.184]				
Constant	7.385	10.788	15.567	15.146	15.028	15.028	15.028	9.847	14.461
	[2.346]**	[2.032]**	[1.764]**	[1.518]**	[1.858]**	[1.583]**	[1.566]**	[2.506]**	[1.976]**
Observations	1444	1438	1444	1444	1444	1444	1444	1444	1444
R-squared	0.15	0.15	0.15	0.14	0.16	0.16	0.16	0.16	0.16
System test: Chi2						893.1	87.4		
df						100	25		
P value:						0.000	0.000		
adjusted coeff:			0.293						

Standard errors in parentheses

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

Standard errors in the LW regression have been bootstrapped, with 200 replications

Table 5: The determinants of body mass in the DHS

Dependent variable: BMI	[1]	[2]	[3]	[4]	[5]	[6]
	PCA	Proxies tel	Proxies all	LW	LW2	PCA2
employed	0.209	0.337	0.272	0.272	0.272	0.252
	[0.127]	[0.127]**	[0.127]*	[0.122]*	[0.119]*	[0.127]*
age	0.411	0.427	0.409	0.409	0.409	0.406
	[0.020]**	[0.020]**	[0.020]**	[0.020]**	[0.021]**	[0.020]**
age squared	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
	[0.000]**	[0.000]**	[0.000]**	[0.000]**	[0.000]**	[0.000]**
years education	0.069	0.099	0.056	0.056	0.054	0.057
	[0.018]**	[0.017]**	[0.018]**	[0.019]**	[0.019]**	[0.018]**
Asset index/ asset 1	0.478	0.303	0.105	0.657	0.259	0.338
	[0.044]**	[0.043]**	[0.047]*	[0.067]**	[0.059]**	[0.105]**
smoker	-2.132	-2.212	-2.102	-2.102	-2.102	-2.097
	[0.139]**	[0.140]**	[0.139]**	[0.130]**	[0.135]**	[0.140]**
female	2.957	2.942	2.971	2.971	2.972	2.975
	[0.127]**	[0.128]**	[0.127]**	[0.128]**	[0.127]**	[0.127]**
children	0.16	0.163	0.155	0.155	0.155	0.154
	[0.033]**	[0.033]**	[0.033]**	[0.036]**	[0.037]**	[0.033]**
number of adults	-0.081	-0.03	-0.104	-0.104	-0.104	-0.095
	[0.038]*	[0.038]	[0.038]**	[0.041]*	[0.041]*	[0.038]*
indian	-2.519	-2.119	-2.607	-2.607	-2.662	-2.615
	[0.319]**	[0.316]**	[0.324]**	[0.320]**	[0.303]**	[0.322]**
white	-1.557	-0.722	-1.401	-1.401	-1.374	-1.502
	[0.260]**	[0.240]**	[0.294]**	[0.279]**	[0.263]**	[0.281]**
coloured	-0.66	-0.326	-0.696	-0.696	-0.541	-0.593
	[0.176]**	[0.172]+	[0.185]**	[0.192]**	[0.180]**	[0.176]**
urban	0.641	0.92	0.764	0.764	0.713	0.777
	[0.135]**	[0.131]**	[0.145]**	[0.140]**	[0.144]**	[0.144]**
Electricity			-0.023		-0.025	-0.252
			[0.161]		[0.148]	[0.181]
Television			0.647		0.716	0.474
			[0.159]**		[0.158]**	[0.183]**

Refrigerator			0.486		0.471	0.295
			[0.171]**		[0.177]**	[0.202]
Bicycle			-0.311		-0.313	-0.539
			[0.158]*		[0.166]+	[0.173]**
Car			0.839		0.794	0.59
			[0.174]**		[0.186]**	[0.214]**
Sheep/cattle			0.368		0.372	0.501
			[0.197]+		[0.172]*	[0.194]**
Radio			0.451			
			[0.155]**			
Pesonal Computer (PC)			-1.095			
			[0.302]**			
Washing Machine			0.77			
			[0.209]**			
Motorcycle			0.24			
			[0.439]			
Donkey/horse			-0.187			
			[0.337]			
Constant	14.203	12.801	13	13	12.995	14.001
	[0.528]**	[0.504]**	[0.516]**	[0.500]**	[0.527]**	[0.584]**
Observations	10299	10299	10299	10299	10299	10299
R-squared	0.19	0.19	0.20	0.20	0.20	0.19
System test: Chi2				5386.4	76.3	
df				132	24	
P value:				0.000	0.000	

Standard errors in parentheses

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

Standard errors in the LW regression have been bootstrapped, with 200 replications

About DatatFirst

DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
- research to improve the quality of African survey data
- training of African data managers for better data curation on the continent
- training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708