DataFirst Technical Papers

DataFirst

SALDRU

Estimating expenditure impacts without expenditure data using asset proxies

by

Martin Wittenberg

Technical Paper Series

Number 10

About the Author(s) and Acknowledgments

Martin Wittenberg is a Professor in the School of Economics at the University of Cape Town, a SALDRU Associate and one of the principal investigators of the Mellon Data Quality Project. For more information about the project visit http://www.datafirst.uct.ac.za.

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

Recommended citation

Wittenberg, M. (2009). Estimating expenditure impacts without expenditure data using asset proxies. A DataFirst Technical Paper Number 10. 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

Estimating expenditure impacts without expenditure data using asset proxies

Martin Wittenberg
School of Economics and SALDRU
University of Cape Town

February 11, 2009

Abstract

When asset indices are used in regressions the coefficients obtained are typically difficult to interpret. We show how lower bounds on expenditure effects can be extracted, if the relationship between the assets and expenditure can be calibrated on an auxiliary data set.

Key words: asset index, proxy variables, attenuation, obesity

JEL codes: C13, C81

Ever since Filmer and Pritchett (2001) suggested how one might "estimate wealth effects without expenditure data – or tears" many authors have utilised asset indices as proxies for expenditure or "socio-economic status" in multiple regressions. Applications include the study of childhood cognitive development (Paxson and Schady 2007), the situation of orphans (Ainsworth and Filmer 2006) and the impact of economic development on child health (Boyle et al. 2006). Extensions to the technique have been proposed by McKenzie (2005) and reviews of the utility of the procedure are provided *inter alia* by Bollen, Glanville and Stecklov (2002) and Montgomery, Gragnolati, Burke and Paredes (2000).

One of the key difficulties facing researchers in using these indices is that the regression coefficients are hard to interpret. In this paper we will suggest that it is possible to extract economically meaningful interpretations provided that one is able to relate the asset index to expenditure in an auxiliary survey.

Our empirical application is the economic determinants of obesity. The spread of obesity internationally has been linked to a "nutrition transition" (Popkin 1999) in which economic factors play an important part (Chou, Grossman and Saffer 2002, Philipson and Posner 2003). Obesity in turn has important medical and social impacts. 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 would therefore be useful. Unfortunately, as Filmer and Pritchett (2001) noted, the Demographic and Health Surveys (DHSs), the largest available data sets with anthropometric information, do not have adequate socio-economic information. We are therefore forced to rely on asset variables to proxy for household expenditure.

1 The model

We assume that the asset variables function purely as proxies for the missing expenditure information. In particular we assume that the k asset variables can all be written in the form

$$\begin{aligned} \mathbf{a}_1 &= \rho_1 \mathbf{z} + \boldsymbol{\nu}_1 \\ \mathbf{a}_2 &= \rho_2 \mathbf{z} + \boldsymbol{\nu}_2 \\ &\vdots \\ \mathbf{a}_k &= \rho_k \mathbf{z} + \boldsymbol{\nu}_k \end{aligned} \tag{1}$$

where $\mathbf{z}$ is the latent variable of interest (household expenditure) and the $\boldsymbol{\nu}_j$ terms are idiosyncratic errors. From this it follows that any asset index $\mathbf{a}_{index}$ created as the linear combination of the asset variables (such as the first principal component) can likewise be written in this form, i.e.

$$\mathbf{a}_{index} = \rho_{index} \mathbf{z} + \boldsymbol{\xi} \tag{2}$$

We assume that the regression of interest can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \boldsymbol{\varepsilon} \tag{3}$$

where γ is the parameter of interest, but $\mathbf{z}$ is not available in our data set. Note that if we use $\mathbf{a}_{index}$ to proxy for $\mathbf{z}$ our structural regression becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a}_{index}\left(\frac{\gamma}{\rho_{index}}\right) + \boldsymbol{\eta}$$

where $\boldsymbol{\eta} = \boldsymbol{\varepsilon} - \frac{1}{\rho_{index}}\boldsymbol{\xi}$. Estimating this equation by OLS we will get an attenuated estimate of $\frac{\gamma}{\rho_{index}}$ due to the correlation between $\mathbf{a}_{index}$ and the regression error $\boldsymbol{\eta}$. Nevertheless this estimate would give us a lower bound on the absolute value of $\frac{\gamma}{\rho_{index}}$, which would be informative about the parameter of interest, provided that we knew what ρ_{index} was.

Our approach is quite simple. We project $\mathbf{a}_{index}$ on $\mathbf{z}$ in a data set where we have both of them available. We then rescale the asset index as $\mathbf{a}^* = \frac{1}{\widehat{\rho}_{index}}\mathbf{a}_{index}$ and use this in our regression, which becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a}^*\gamma + \boldsymbol{\eta} \tag{4}$$

The estimate we obtain in this way will provide a lower bound on the absolute value of γ. It is therefore interpretable directly as an expenditure effect.

In fact we can improve on this lower bound, using a result of Lubotsky and Wittenberg (2006). They show that the attenuation error can be reduced by including all proxies in the multiple regression and aggregating up their coefficients as $\boldsymbol{\rho}'\mathbf{b}$ where $\mathbf{b}$ is the vector of coefficients of the asset proxies. A simple procedure is to project each of the asset proxies on $\mathbf{z}$ on the auxiliary data set and to create the rescaled indices

$$\begin{aligned} \mathbf{a}_1^* &= \frac{1}{\widehat{\rho}_1}\mathbf{a}_1 \\ &\vdots \\ \mathbf{a}_k^* &= \frac{1}{\widehat{\rho}_k}\mathbf{a}_k \end{aligned}$$

and then to estimate the proxy regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^{k}\mathbf{a}_j^* b_j + \boldsymbol{\eta} \tag{5}$$

where $\boldsymbol{\eta} = \boldsymbol{\varepsilon} - \sum_j \frac{1}{\widehat{\rho}_j}\widehat{\boldsymbol{\nu}}_j$ and $\widehat{\boldsymbol{\nu}}_j$ are the regression residuals from the k proxy equations. A lower bound on γ can then be estimated as

$$\widehat{\gamma} = \sum_{j=1}^{k}\widehat{b}_j \tag{6}$$

In our data sets this procedure improves on the estimate obtained by the composite index by between thirteen and twenty-seven percent.

Both procedures have the disadvantage that we need to have precisely the same assets available in the auxiliary data set as in the main data set under consideration. In our case, however, there are a few asset variables that are available in the DHS that are not available in the Income and Expenditure Survey, our auxiliary data set. In this case we resort to a third variant. We project one of the assets (say $\mathbf{a}_1$) that **is** available in both data sets on $\mathbf{z}$ and rescale it as before. We then use this rescaled asset $\mathbf{a}_1^*$ to estimate $\widehat{\rho}_2, \; , \widehat{\rho}_k$ within the main data set by the

GMM procedure as outlined in the Lubotsky and Wittenberg (2006) article or in Wittenberg (2007). The proxy regression in this case is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a}_1^* b_1 + \sum_{j=2}^{k} \mathbf{a}_j b_j + \boldsymbol{\eta}$$

We can now estimate a lower bound on γ as

$$\widehat{\gamma} = \widehat{b}_1 + \sum_{j=2}^{k} \widehat{\rho}_j \widehat{b}_j \tag{7}$$

In this case we need only one asset variable that is common to both data sets, although it should be one that is well correlated with expenditure.

2 The data sets and estimation issues

The dependent variable in our regressions is the body mass index, defined as weight (in kilograms) divided by height (in metres) squared. A person with a BMI in excess of twenty-five is defined as overweight, while thirty is the threshold for being classified as obese. We estimate the relationship between the body mass index and expenditure on two data sets. The first is the 1998 KwaZulu-Natal Income Dynamics Survey (KIDS), which was concentrated in the KwaZulu Natal province. This is a relatively small survey that has both asset information as well as reasonably good socio-economic data. However, it is limited by sample size and its limited geographic coverage. We use this data set to see how our procedure performs when we do have a good measure of household expenditure available.

The second survey is the South African Demographic and Health Survey from 1998. This is the only nationally representative survey that has anthropometric information. It also has a large sample size. Like all Demographic and Health Surveys, its socio-economic information is seriously deficient. We calibrate our asset indices on the Income and Expenditure Survey (IES) of September 2000, which was matched with the Labour Force Survey of that period.

In our analysis of the DHS we have no information about what the "true" coefficients should be, but we can compare our estimates from the DHS to the estimates which were obtained from the smaller data set. It turns out that the coefficients that we get when we analyse the DHS are similar to those that we obtained on the KIDS data, which increases our confidence that the procedure works. In order to make this comparison we have used broadly similar control variables in each of the data sets.

One important issue that needs to be addressed is how to obtain appropriate standard errors for our estimates, given that the rescaling of the asset variables makes these stochastic. In the case of the KIDS data we bootstrapped the entire procedure, from the initial rescaling to the final estimation of the lower bound on the expenditure effect.

In the case of the national datasets, the estimates of the rescaling coefficients were obtained from the IES. These stochastic $\widehat{\rho}_j$ variables would therefore have been independent of the asset variables $\mathbf{a}_j$ used in the DHS. Consequently for each bootstrap sample from the DHS, we took a random draw $\widehat{\rho}_j^*$ from a normal distribution with mean $\widehat{\rho}_j$ and variance $\widehat{\sigma}^2_{\rho_j}$ as obtained from the regression on the IES, and rescaled those bootstrap sample asset proxies as $\frac{1}{\rho_j^*}\mathbf{a}_j$. We then ran the regression (equation 5) and calculated the estimate (6) on that bootstrap sample. In all cases two hundred replications were done.

3 Results

Our empirical results are contained in Table 1. The first three columns present the results from the KIDS survey. In column 1 we have used the log of household expenditure as our explanatory variable. To calibrate the meaning of the coefficient of 1.309 it is useful to note that average height among adults in South Africa (as measured in the DHS) was 1.62m (5 ft 3 inches), suggesting that a one unit change in the log of household expenditure would lead to an increase in weight of 3. 44 kg (7.5 lb) for a person of average height. A shift from the 25th to the 75th percentile of the South African income distribution, i.e. 1.4 units on the log scale, imply a weight difference of 4.8 kg (10. 5 lb) for individuals of average height.

In the second column we have used a rescaled Filmer-Pritchett style asset index. The unscaled coefficient was 0.484. The rescaling increases the coefficient, but this lower bound is still less than half the "true" coefficient.. The theoretically more efficient estimate in column 3, calculated according to equation 6, significantly improves on that asset, although it still shows attenuation of around 40%. Nevertheless it still suggests a meaningful impact both in the statistical as well as real sense. Individuals at the 75th percentile of South Africa's income distribution would be expected to be 2. 87 kg (6. 3 lb) heavier on average than individuals at the 25th percentile. This lower bound on the true impact provides more useful information than comparing individuals at the 25th and 75th percentile of the asset distribution, since it is not generally clear how that might translate into "real" well-being. Furthermore the asset distribution is itself sensitive to the assets used in the construction of the index.

In the final three columns we implement our estimation procedure on the DHS. In column four we rescale an asset index constructed within the DHS according to the relationship between the "same" asset index calculated on the IES and log expenditure measured on the IES. The point estimate is very close to that obtained by principal components in the KIDS survey. Putting the assets into the regression and aggregating them up according to equation 6 produces a higher estimate (column five). The highest "lower bound" is obtained when we use more asset variables available within the DHS and estimate the ρ vector using the correlation structure within the DHS (column six). The coefficients are aggregated up according to equation 7. This estimate is again close to the equivalent one on the KIDS data set.

Given the results of the KIDS survey, we would expect the "true" coefficient to be substantially larger than the coefficients estimated in columns four to six, but the procedures seem to produce a meaningful and valid lower bound on that parameter. The important point is that we have no other way of fixing the true parameter, since the KIDS survey is not nationally representative.

4 Conclusion

In many cases we are interested in estimating the real impacts of log expenditure on a particular response variable. In cases where we do not have that information available but have a set of asset proxies we can obtain lower bounds on those impacts by calibrating these variables against log expenditure in an auxiliary data set. Testing this procedure on a data set where we have both suggests that even where there is significant attenuation we can obtain meaningful information. Applying this procedure to the case of obesity in South Africa suggests that increases in household income are associated with statistically significant and economically meaningful increases in weight.

References

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

Bollen, Kenneth A., Jennifer L. Glanville, and Guy Stecklov, "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.

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.

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.

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., 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.

Philipson, Tomas J. and Richard A. Posner, "The long-run growth in obesity as a function of technological change," *Perspectives in Biology and Medicine*, Summer 2003, *46* (3).

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

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.

Table 1: Determinants of the body mass index in two surveys

	KIDS			DHS		
	[1]	[2]	[3]	[4]	[5]	[6]
	ln Exp	PCA$^{(a)}$	LW$^{(a)}$	PCA$^{(b)}$	LW$^{(b)}$	LW$^{(c)}$
Expenditure / proxy	1.309	0.614	0.781	0.600	0.680	0.758
	[0.280]**	[0.162]**	[0.191]**	[0.054]**	[0.055]**	[0.074]**
employed	0.279	0.208	0.244	0.191	0.243	0.236
	[0.376]	[0.342]	[0.364]	[0.123]	[0.126]+	[0.129]+
age	0.36	0.381	0.362	0.389	0.386	0.381
	[0.066]**	[0.058]**	[0.054]**	[0.020]**	[0.018]**	[0.021]**
age^2	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003
	[0.001]**	[0.001]**	[0.001]**	[0.000]**	[0.000]**	[0.000]**
educ	0.005	0.037	0.005	0.111	0.101	0.088
	[0.059]	[0.054]	[0.056]	[0.017]**	[0.017]**	[0.018]**
adults	-0.042	0.030	0.025	-0.087	-0.109	-0.129
	[0.088]	[0.096]	[0.083]	[0.037]*	[0.038]**	[0.040]**
children	0.119	0.211	0.172	0.193	0.176	0.181
	[0.072]+	[0.081]**	[0.075]*	[0.033]**	[0.033]**	[0.036]**
female	4.032	4.022	4.04	3.737	3.755	3.732
	[0.356]**	[0.347]**	[0.325]**	[0.107]**	[0.101]**	[0.113]**
urban	1.107	0.733	0.655	0.496	0.700	0.644
	[0.513]*	[0.592]	[0.593]	[0.141]**	[0.148]**	[0.135]**
city	1.791	1.67	1.575			
	[0.626]**	[0.698]*	[0.699]*			
indian	-4.716	-3.837	-3.602	-2.629	-2.668	-2.900
	[0.644]**	[0.582]**	[0.681]**	[0.326]**	[0.308]**	[0.306]**
white				-1.615	-1.490	-1.900
				[0.262]**	[0.282]**	[0.292]**
coloured				-1.086	-1.062	-1.304
				[0.187]**	[0.180]**	[0.196]**
Obs	1444	1444	1444	10299	10299	10299
R^2	0.15	0.15	0.16	0.17	0.18	0.18

Standard errors in brackets

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

(a) Assets used: telephone, electricity, television, refrigerator, furniture, jewellery, electrical appliance, car, bicycle, cattle, sheep

(b) Assets used: telephone, electricity, television, radio, computer, car, sheep or cattle

(c) Assets used: telephone, electricity, radio, television, refrigerator, computer, washing machine, bicycle, motorcycle, car, donkey or horse, sheep or cattle

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.

DataFirst

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