Attenuation Bias, Measurement Error & Principal Component Analysis

Isaac Liu, Nicolás Martorell & Paul Opheim

May 16, 2021

Abstract

Many variables of interest in Economics are not directly available as empirical data. Instead, economists often use other variables that are imperfect measurements of the true focus of their analysis. These available variables are known as *proxies* or "variables measured with error", and, if they suffer from classical measurement error, their use causes *attenuation bias* when they are used as independent variables in econometric estimation. Traditionally, instrumental variables are used as a shock of exogeneity to get rid of this bias, but finding truly exogenous variables that satisfy the exclusion restriction is difficult, and so this method can often not be feasibly applied.

As an alternative to dealing with attenuation bias, we propose the use of Principal Component Analysis (PCA) over several variables measured with error. When there are multiple observed variables driven by a single "true" one, we propose to use PCA over these variables to extract the "true" variable. We then use this extracted value and use it in a standard OLS regression, thus providing a solution to attenuation bias that does not require the strong assumptions of instrumental variable analysis.

To show the properties and behaviour of our estimator on large samples under standard assumptions, we present a theoretical framework and a Monte-Carlo analysis. Additionally, we explore a basic empirical application to our method, by estimating the effect of economic development on life expectancy at birth. Since there is no consensus on how to measure economic development, we take a sample of different variables that may measure economic development with error (GDP per capita, GNI per capita, Household Income Per Capita, among others) over which we apply PCA to apply our identification strategy.

1 Theoretical framework

Consider a model where the outcome is denoted by y_i. This outcome depends on a variable of interest denoted by t_i and a vector of covariates denoted by $X_i = (x_{i,1}, x_{i,2}, \ldots x_{i,p})'$. Additionally, consider a vector of variables $X_i^* = (x_{i,1}^*, x_{i,2}^*, \ldots x_{i,p}^*)'$ that correspond to the covariates X_i but observed with measurement error, where $x_{i,k}^* = x_{i,k} + \eta_{i,k}$ with $\eta_{i,k} \sim iid(0, \sigma_{\eta_k}^2)$, $\mathrm{E}(x_{i,k}'\eta_{i,k}) = 0, \forall i$, $\mathrm{E}(x_{i,k}'\eta_{j,l}) = 0, \forall i \neq j$ and $k \neq l$, and $\mathrm{E}(\eta_{i,k}'\eta_{j,l}) = 0, \forall i \neq j$ and $k \neq l$. Therefore, each $x_{i,k}^*$ suffers classical measurement error. Note that $\mathrm{E}(x_{i,k}) = \mathrm{E}(x_{i,k}^*) = \mu_{x_k}$ and that $\mathrm{V}(x_{i,k}) = \sigma_{x_k}^2$ while $\mathrm{V}(x_{i,k}^*) = \sigma_{x_k}^2 + \sigma_{\eta_k}^2 \geq \sigma_{x_k}^2$.

1.1 Data Generation Process

Assume that the outcome y_i is determined by the following Data Generation Process (DGP):

$$y_i = \gamma t_i + X_i'\beta + \epsilon_i \tag{1}$$

where γ is the parameter of the variable of interest t_i, $\beta = (\beta_1, \beta_2, \ldots \beta_p)'$ is the vector of the parameters of the covariates X_i including a constant and $\epsilon_i \sim \mathrm{iid}(0, \sigma_\epsilon^2)$. Under this specification, the coefficients are such that:

$$\begin{pmatrix} \gamma \\ \beta \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{yX} \end{pmatrix} \tag{2}$$

Suppose that the econometrician has access to t_i but, instead of X_i she observes X_i^*. Then, she specifies the following linear model

$$y_i = \gamma^* t_i + X_i^{*\prime}\beta^* + \zeta_i \tag{3}$$

the coefficients would be such that

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \Sigma_{tX^*} \\ \Sigma_{X^*t} & \Sigma_{X^*} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{yX^*} \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X + \Sigma_\eta \end{pmatrix}^{-1} \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{5}$$

To the see the implications of the of this measurement error in the covariates, consider a simple case where the DGP depends only of the variable of interest and a covariate such that:

$$\begin{pmatrix} \gamma \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{6}$$

and with $\sigma_t^2 = \Sigma_X = \Sigma_\eta = 1$ while $\Sigma_{Xt} = 0.6$. Then

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$
$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} 1.37 \\ 0.39 \end{pmatrix}$$

Clearly, both coefficients shows bias when the econometrician assumes a DGP with X_i^*: while there is attenuation bias on the coefficient of the covariate, the coefficient of the variable of interest is biased upward given that both variable have positive correlation.

1.2 Principal Component Regression as bias-correction method

The classical solution for the measurement-error induced bias in econometrics has been the usage of instrumental variables. Suppose an instrument Z_i that satisfies the relevance condition $\mathrm{E}(Z_i'X_i) \neq 0$ and $\mathrm{E}(Z_i't_i) \neq 0$, and also the exclusion restriction $\mathrm{E}(Z_i'\epsilon_i) = \mathrm{E}(Z_i'\zeta_i) = \mathrm{E}(Z_i'\eta_{i,k}) = 0$, for all i and k. Then premultiplying by Z_i we have

$$Z_i'y_i = Z_i'\gamma^* t_i + Z_i'X_i^{*\prime}\beta^* + Z_i'\zeta_i \tag{7}$$

and so

$$\begin{pmatrix} \gamma^{IV} \\ \beta^{IV} \end{pmatrix} = \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} + \Sigma_{Z\eta} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{8}$$

$$= \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{9}$$

$$\begin{pmatrix} \gamma^{IV} \\ \beta^{IV} \end{pmatrix} = \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{10}$$

However, finding reliable source of exogeneity is difficult, as is proving the exclusion condition. Therefore, the use of IV as a bias-correction method should be taken with care given that its feasibilty is hard.

Alternatively, we propose an alternative bias-correction method when there are several miss-measured variables for each covariate, that is when we have more than one $x_{i,k}^*$ for every $x_{i,k}$. Given that in all the miss-measured variables the underlying value is the real value, one could think of extracting the underlying true $x_{i,k}$ through a linear combination of the different $x_{i,k}^*$. Then, we could treat all the $x_{i,k}^*$ as variables that share compononents as follows

$$h_j = \underset{h'h=1, h'h_1=0, \ldots, h'h_{j-1}=0}{\operatorname{argmax}} \operatorname{var}\left[h'X_k^*\right] \tag{11}$$

where h_j is the eigenvector of Σ associated with the j^{th} ordered eigenvalue λ_j of $\Sigma_{X_k^*}$, and the principal components of X_k^* are $U_j = h_j'X_k^*$, where h_j is the eigenvector of Σ associated with the j^{th} ordered eigenvalue λ_j of Σ.

Under our assumptions, the vector of missmeasured values X_k^* of $x_{i,k}$, share only one principal component which is precisely $x_{i,k}$. Then, we only have one principal component, $x_{i,k}$, and so the $x_{i,k}$ is such that

$$x_{i,k} = h_k'X_k^* \tag{12}$$

Finally, we could then retrieve the vector of true variables X_i

$$X_i = HX_i^* \tag{13}$$

where H is a matrix such that

$$H = \begin{pmatrix} h_1 & 0 & 0 & \ldots & 0 \\ 0 & h_2 & 0 & \ldots & 0 \\ \vdots & \ddots & h_3 & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots \ddots & h_p \end{pmatrix}$$

and h_k is the vector of eigenvalues for the variable $x_{i,k}$.
Our new linear model then would be

$$y_i = \gamma^{PCR} t_i + H X_i^{*\prime} \beta^{PCR} + \epsilon_i \tag{14}$$

where the coefficients are as follows

$$\begin{pmatrix} \gamma^{PCR} \\ \beta^{PCR} \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{y,HX^*} \end{pmatrix} \tag{15}$$

$$= \begin{pmatrix} \sigma_t^2 & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{16}$$

$$= \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{17}$$

where the last equality comes from (13).

Application: Government Share of Healthcare Spending and Life Expectancy

In the left column in the table below I first regress the life expectancy at birth for all individuals in a given country and year on a measure of government spending as a share of total health expenditure. In the middle column I include the economic controls/covariates of GDP per capita (PPP), GNI per capita (PPP), Survey Mean Income/Consumption Per Capita, ILO GDP per person employed, and Net Foreign Assets Per Capita, all from the World Bank. In the rightmost column I instead use the first principal component combining these covariates.

I standardize all variables by subtracting the mean and dividing by the standard deviation, linearly interpolate data between known observations, and remove country-years with missing values for any of the economic indicators.

	Life Expectancy at Birth (Years)				
	(1)	(2)	(3)	(4)	(5)
Govt. Share of Health Exp.	0.613***	0.308***	-0.016	0.343***	-0.012
	(0.019)	(0.020)	(0.019)	(0.019)	(0.018)
Covariates	None	Econ Indicators	Econ Indicators	PCs	PCs
Fixed Effects	No	No	Yes	No	Yes
Observations	1,799	1,799	1,799	1,799	1,799
R^2	0.376	0.586	0.987	0.536	0.987
Adjusted R^2	0.375	0.582	0.985	0.535	0.985
Residual Std. Error	0.791	0.646	0.122	0.682	0.123
F Statistic	1081.530***	168.157***	922.656***	1036.417***	1114.735***

Note: *p<0.1; **p<0.05; ***p<0.01
All variables are standardized

With Ginis instead of econ controls

	Life Expectancy at Birth (Years)				
	(1)	(2)	(3)	(4)	(5)
Govt. Share of Health Exp.	0.697***	0.652***	0.028	0.692***	0.026
	(0.034)	(0.040)	(0.024)	(0.037)	(0.023)
Covariates	None	Ginis	Ginis	PCs	PCs
Fixed Effects	No	No	Yes	No	Yes
Observations	322	322	322	322	322
R^2	0.566	0.596	0.999	0.566	0.999
Adjusted R^2	0.565	0.583	0.998	0.564	0.998
Residual Std. Error	0.596	0.583	0.040	0.597	0.040
F Statistic	417.382***	45.905***	1421.063***	208.238***	5197.564***

Note: *p<0.1; **p<0.05; ***p<0.01
All variables are standardized.

Health share of gdp

	Life Expectancy at Birth (Years)				
	(1)	(2)	(3)	(4)	(5)
GDP Share of Health Exp.	0.258***	0.184***	-0.024	0.144***	-0.020
	(0.023)	(0.019)	(0.031)	(0.017)	(0.030)
Covariates	None	Econ Indicators	Econ Indicators	PCs	PCs
Fixed Effects	No	No	Yes	No	Yes
Observations	1,799	1,799	1,799	1,799	1,799
R^2	0.067	0.553	0.987	0.467	0.987
Adjusted R^2	0.066	0.549	0.985	0.467	0.985
Residual Std. Error	0.967	0.672	0.122	0.730	0.123
F Statistic	128.396***	146.946***	772.041***	787.819***	384.737***

Note: *p<0.1; **p<0.05; ***p<0.01
All variables are standardized.

Appendix

Figure 1: Correlations Between Covariates and Life Expectancy

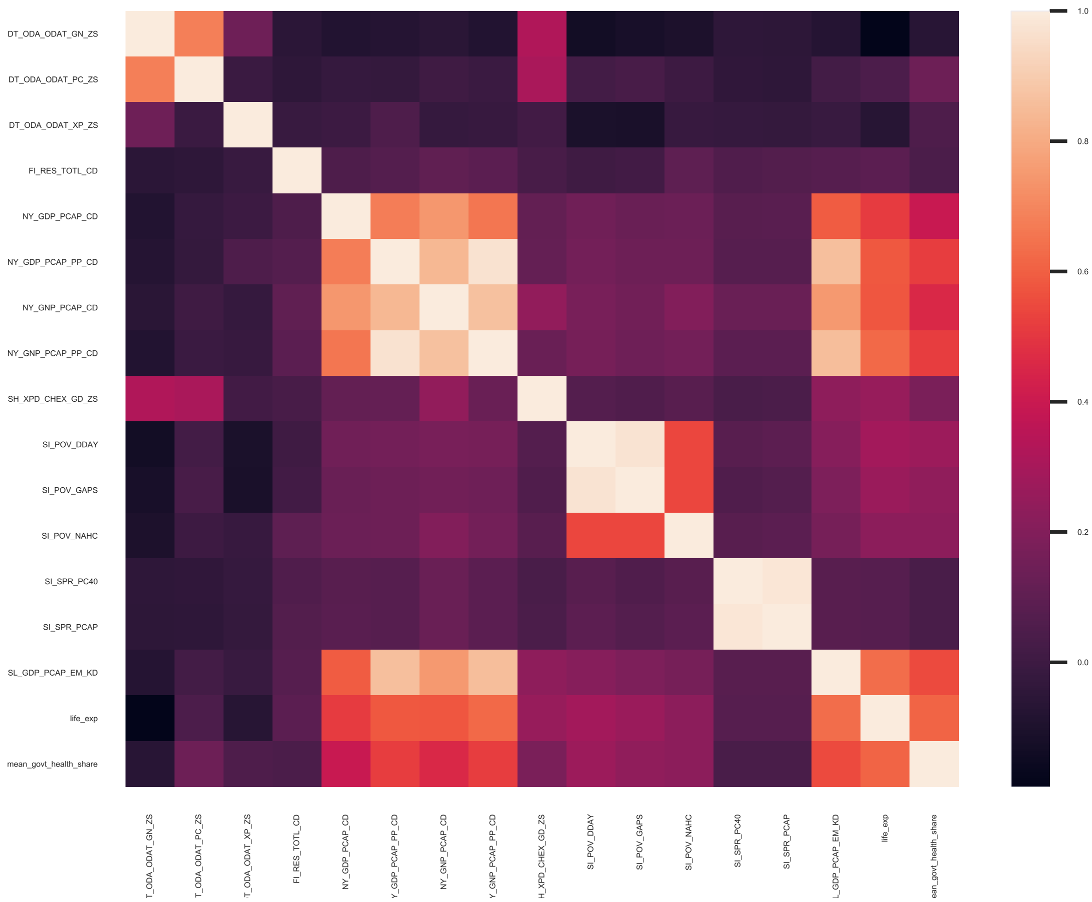

Figure 2: Economic Measures PCA Loadings

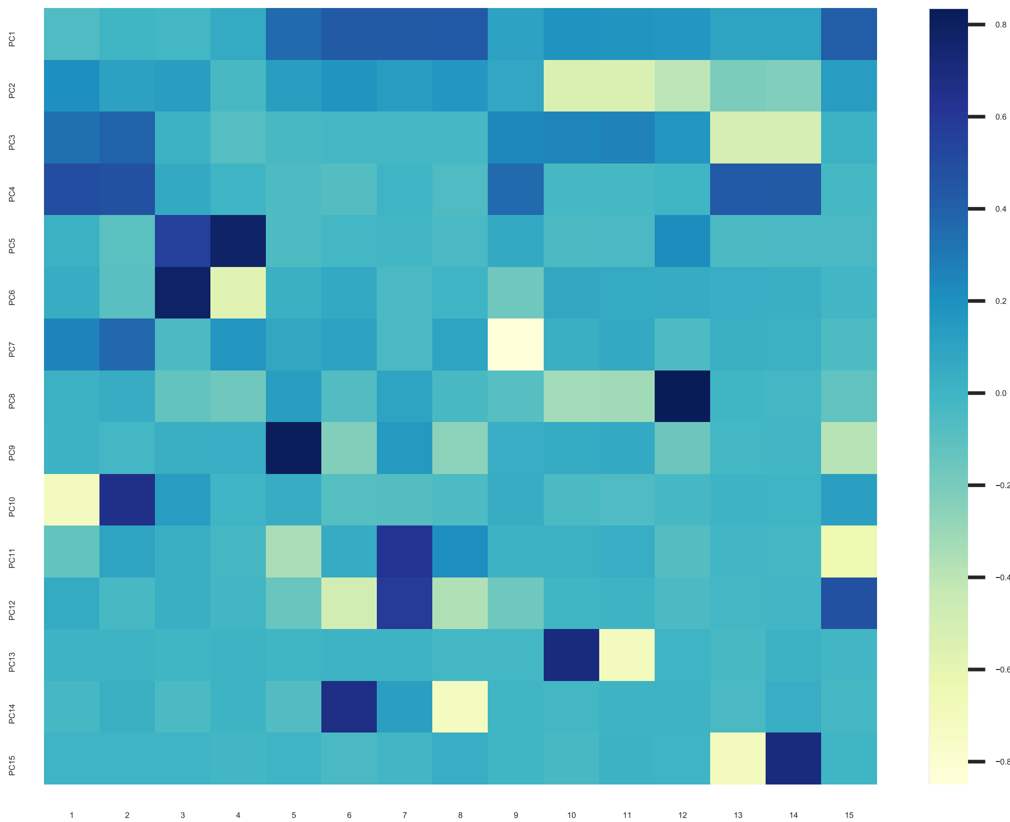

Figure 3: Economic Measures PCA Share of Variance Explained