Principal Component Regression as a Solution to Measurement Error Bias

Isaac Liu, Nicolás Martorell & Paul Opheim

June 3, 2021

Abstract

We show that using Principal Component Regression (PCR) is a useful solution to bias introduced by measurement error in auxilliary covariates included in a regression. We show its usefulness through econometric theory and then use Monte Carlo simulations to show how it provides benefits for different parameters and correlations between the true covariate and the variable of interest. We also find that PCR is outperformed by an instrumental variables approach. We then apply these methods to study the relationship between life expectancy and the level of government involvement in a country's healthcare system.

Introduction

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, are relevant for the model specification, and are correlated with the variable of interest, their use biases the coefficient of the variable of interest even if it does not suffer of measurement error. Traditionally, instrumental variables are used to get rid of measurement error induced bias.

As an alternative method of dealing with this problem, 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. One may then use this extracted value in a standard OLS regression (often referred to as PCR or Principal Components Regression), thus providing a way to identify the parameter of interest that does not require the assumptions of instrumental variable analysis. The method also allows for more complex and possibly more optimal weightings of mismeasurements relative to simple averaging, and is less vulnerable to the curse of dimensionality relative to the inclusion of many covariates.

This estimator ties into earlier literature considering the intersection of factor models and principal components analysis and measurement error and latent variables problems. Somewhat similarly to our methods, Nagasawa (2020) develops the use of a proxy variable to deal with unobserved heterogeneity in nuisance parameters and uses a partial effects method. Differing from our setting, Schennach (2016) focuses on nonclassical measurement error and nonlinear cases and notes the usefulness of factor methods and some cases where they are of more use than instrumental variables. Wegge (1996) considers a setting in which measurement error regression models are factor analysis models, with the correct regressors being the factors. Latent factors are uncorrelated with the errors. Focusing on measurement error in the main regressor, Schofield (2015) combines solutions from structural equations modelling and item response theory to deal with misestimation. Finally, Heckman, Schennach and Williams (2010) considers a situation similar to ours, except involving matching estimators. Without correction, matching estimators can be harmed by mismeasured conditioning variables. However, average treatment effects can be identified using factor proxies, and without need for normalization.

In this paper, we present a theoretical framework and a Monte-Carlo analysis in order to show the properties and behavior of our estimator on large samples under standard assumptions. Additionally, we explore a basic empirical application of our method, by estimating the relationship between 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, Income per Employed Person, among others) over which we apply PCA to estimate coefficients. Our estimator generally behaves as expected in this empirical setting, though it is unclear whether it performs any better or worse than the direct inclusion of covariates, their averaging, or the instrumentation of mismeasured variables with each other.

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 $X_i = (x_{i,1}, x_{i,2}, \dots x_{i,p})'$. Additionally, consider a vector of variables $X_i^* = (x_{i,1}^*, x_{i,2}^*, \dots x_{i,p}^*)'$ that correspond to the covariates X_i but observed with measurement error, where $x_{i,k}^* = x_{i,k} + \eta_{i,k}$ with $\eta_{i,k} \sim iid(0, \sigma^2_{\eta_k})$, $\mathrm{E}(x'_{i,k}\eta_{i,k}) = 0, \forall i$, $\mathrm{E}(x'_{i,k}\eta_{j,l}) = 0, \forall i \neq j$ 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 $\mathrm{V}(x_{i,k}) = \sigma^2_{x_k}$ while $\mathrm{V}(x^*_{i,k}) = \sigma^2_{x_k} + \sigma^2_{\eta_k} \geq \sigma^2_{x_k}$.

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 $\epsilon_i \sim \text{iid}(0, \sigma^2_\epsilon)$. Under this specification, the coefficients are such that:

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

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:

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

The coefficients would be such that:

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} \sigma^2_t & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X + \Sigma_\eta \end{pmatrix}^{-1} \begin{pmatrix} \sigma^2_t & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{4}$$

Without loss of generality, assume that the abovementioned GDP consists only of two variables as follows:

$$y_i = \gamma^* t_i + \beta^* x_i * + \zeta_i \tag{5}$$

Then the coefficient of our variable of interest will be biased.

Claim 1 *Under measurement error in the covariates, the coefficient is such that:*

$$\gamma^* = \gamma + \beta \frac{\text{cov}(t,x)(\sigma^2_{x^*} - \sigma^2_x)}{\sigma^2_t \sigma^2_{x^*} - \text{cov}(t,x)^2} \tag{6}$$

Proof of claim 1.1 *See Theory Appendix.*

From (10) it is clear that when $\text{cov}(t,x) \neq 0$ and that x is measured with error (i.e $\sigma^2_{x^*} > \sigma^2_x$), the coefficient of our variable of interest is biased. If t and x are independent, then measurement error in x^* does not cause any bias. If there is no measurement error in x^*, then $\sigma^2_{x^*} = \sigma^2_x$ and so we would not face any kind of bias, as one would expect.

Equation (10) also allows us to know the direction of the bias. Given that we are facing measurement error in the covariate, $\sigma^2_{x^*} > \sigma^2_x$ which implies $\sigma^2_{x^*} - \sigma^2_x > 0$. Also, it follows from the *Cauchy-Schwarz* inequality that the denominator is also positive. Then, the direction of the bias will depend on the sign of β and the covariance of t and x, as Table 1 illustrates.

Table 1: Direction of the Bias due to Measurement Error in the Covariate

	$\beta > 0$	$\beta < 0$
$\text{cov}(t, x) > 0$	upward-biased	downward-biased
$\text{cov}(t, x) < 0$	downward-biased	upward-biased

Principal Component Regression as a Bias Correction Method

The classical solution for the measurement-error induced bias in econometrics has been the usage of instrumental variables. Suppose we use as an instrument Z_i another measure of X_i so that

$$Z_i = X_i + \omega_i \tag{7}$$

where $\text{E}(\omega_i) = 0$, $\text{Cov}(\epsilon_i, \omega_i) = 0$ and that ω_i brings new information so that $\text{Cov}(\eta_i, \omega_i) = 0$. Under these conditions, Z_i is a valid instrument.

Claim 2 *Suppose $Z_i = X_i + \omega_i$. If $\text{E}(\omega_i) = 0$, $\text{Cov}(\epsilon_i, \omega_i) = 0$ and $\text{Cov}(\eta_i, \omega_i) = 0$. Then*

$$\text{E}(Z_i \zeta_i) = 0 \textit{ and } \text{E}(Z_i X_i) \neq 0 \tag{8}$$

And so γ can be identified through IV regression.

Proof of claim 2.1 *See Theory Appendix.*

Alternatively, we propose an alternative bias-correction method when there are several mismeasured 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 mismeasured 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 components as follows:

$$h_j = \underset{h'h=1, h'h_1=0, \ldots, h'h_{j-1}=0}{\text{argmax}} \text{var}\left[h' X^*_k\right] \tag{9}$$

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. Then, we could then retrieve the vector of true variables X_i.

Claim 3

$$X_i = H X^*_i \tag{10}$$

where H is a matrix compound of the h_k vectors of eigenvalues of $x_{i,k}$, $\forall i, k$.

Proof of claim 3.1 *See Theory Appendix.*

Our new linear model then would be:

$$y_i = \gamma^{PCR} t_i + H X^{*\prime}_i \beta^{PCR} + \epsilon_i \tag{11}$$

where γ is identified.

Claim 4 *Consider equation* (11). *Then*

$$\gamma^{PCR} = \gamma \tag{12}$$

Proof of claim 4.1 *See Theory Appendix.*

Note that according to equation (17), the true variable $x_{i,k}$ is a linear combination of the mismeasured variables that the researcher may have, where the weights are such that equation (16) is satisfied. This allows us to think about other linear combinations that could be used as a bias-correction method.

In particular, take the case in which $h_k =' (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$, where h_k is a row vector of dimension $(1 \times J)$, and J is the number of mismeasured variables for $x_{i,k}$. Then, equation (17) will be

$$\tilde{x_{i,k}} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} x^*_{i,1} \\ x^*_{i,2} \\ \vdots \\ x^*_{i,J} \end{pmatrix} \tag{13}$$

$$= \frac{1}{n} \sum_{j=1}^{J} x^*_{i,j} \tag{14}$$

That is, the average of the mismeasured variables for $x_{i,k}$ is a feasible linear combination that may correct for the mismeasurement bias problem.

Properties of the Estimator: Monte Carlo Simulations

We complement our theoretical analysis by using Monte Carlo Simulation to analyze the effects of using Principal Components Regression as a method of bias correction. For these simulations, we assume that the true DGP for the data is:

$$y_i = \beta_1 x_i + \beta_2 z_i + u_i$$

where x_i and z_i are single variables drawn from $\mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$ and where ρ is the covariance between our main variable of interest x_i and the covariate z_i. The u_i is drawn from a white noise distribution ($\mathcal{N}(0, 1)$) that is uncorrelated with both x_i and z_i. We then assume (as with the theoretical analysis) that z_i is not directly observable and instead the researchers only have access to p measurements $z^*_{i,j}$ where $z^*_{i,j} = z_i + \eta_j$, where η_j is drawn from a white noise distribution $\mathcal{N}(\mathbf{0}, \Sigma)$, where $\mathbf{0}$ is a p-vector, and where Σ is the p by p identity matrix.

In our simulations, we assume default values of $\rho = 0.5$, $\beta_1 = \beta_2 = 1$, and $p = 5$. We then vary each factor while holding the others fixed, and perform 1,000 simulations of the DGP followed by an OLS regression on either the PCA value from the p measurements of the true z_i, or on a single one of the measurements of z_i. For each simulation, we generate 3,000 observations of y_i, x_i, etc. The first two rows of each panel in Table 2 show the results for different values of ρ.

We first note that when $\rho = 0.5$ or 0.9 then the coefficient on the variable of interest is artificially inflated when we use a single mismeasurement as a covariate (on average, for $\rho = 0.5$, it is roughly 1.29 instead of the true value

Table 2: Average Coefficients for Values of ρ (Uncorrelated Measurement Errors)

	ρ Value				
	-0.9	-0.5	0	0.5	0.9
	Coefficient on Main Variable				
PCA	0.54	0.895	1.001	1.105	1.462
	(0.034)	(0.023)	(0.02)	(0.023)	(0.033)
Single Measurement	0.245	0.714	1.001	1.286	1.756
	(0.025)	(0.023)	(0.023)	(0.024)	(0.025)
All Measurements	0.54	0.895	1.001	1.105	1.461
	(0.034)	(0.023)	(0.02)	(0.023)	(0.033)
Average of Measurements	0.54	0.895	1.001	1.105	1.461
	(0.034)	(0.023)	(0.02)	(0.023)	(0.033)
Instrumental Variable	0.995	0.999	1.001	1.0	1.008
	(0.111)	(0.04)	(0.046)	(0.038)	(0.122)
	Mean Absolute Percentage Error				
PCA	46.0%	10.5%	1.6%	10.5%	46.2%
Single Measurement	75.5%	28.6%	1.8%	28.6%	75.6%
All Measurements	46.0%	10.5%	1.6%	10.5%	46.1%
Average of Measurements	46.0%	10.5%	1.6%	10.5%	46.1%
Instrumental Variable	8.4%	3.0%	2.7%	3.0%	8.6%
Simulations	1,000	1,000	1,000	1,000	1,000

of 1). Conversely, when $\rho = -0.5$ or -0.9 then the coefficient is artificially deflated. Using the PCA value as the covariate reduces this bias for both directions, and brings the main coefficient closer to its true value of 1.0. These results are consistent with our theoretical section, where we argued that a positive covariance between the main variable and the true covariate will lead to an inflation on the main coefficient, while a negative covariance will lead to a deflation of the coefficient. Separately, there is no bias when $\rho = 0$, as predicted. Since there no bias to correct, we do not see gains from using the PCA covariate method for that particular ρ value. These simulation results suggest that using PCR is more effective than a single mismeasured covariate, although there are no gains to using it when the covariance between the covariate and the main variable of interest is close to 0. The simulation appendix contains charts that show that this increase in performance is also true for different values of p, β_1 and β_2.

However, the performance advantages that we see from using PCR could be driven by the benefit of having multiple measurements of our true covariate of interest, as opposed to any special advantages from PCR specifically. We test this question by comparing the estimated β_1 in our PCR regressions with the estimated β_1 when we include all p measurements as separate covariates in the regression, and the β_1 obtained when the covariate is the mean of all p measurements of the true covariate. We also show the results for an instrumental variables regression where we use other measurements of the true covariate as an instrument for a single measurement of the covariate. The results from these regressions for different values of ρ are shown in the bottom three rows of each panel of Table 2.

As one can see from these results (and results for different values of β_1, β_2, and p in the simulation appendix), there does not seem to be a noticeable difference between using PCR, all measurements, or the average of measurements. However, the instrumental variable regression performs far better than these other methods, correcting for almost all bias and moving the estimated coefficient value close to 1 (although with noticeably larger standard errors than the other methods). Thus, our simulations suggest that, when measurement errors are not correlated with one another, there are major benefits to having multiple measurements of a latent covariate of interest. However, under our first framework PCR does not noticeably improve on two other ways of incorporating these other measurements (taking their average or including all measurements as separate covariates) and performs much worse than using instrumental variables regression with these additional measurements.

We also explore how our results differ under a different framework of measurement error. For the following regressions (and those in the simulation appendix) we relax our assumption that the measurement errors are uncorrelated with one another. This means that we edit the error covariance matrix Σ so that the diagonal values continue to be 1 but all other values (that used to be 0) are replaced with 0.5. Since our instrumental variables technique is typically promoted for use with classical measurement error, we wanted to see how the performance of the studied methods changed when that assumption was relaxed. Table 3 shows the results of our five techniques in this new context.

Table 3: Average Coefficients for Values of ρ (Correlated Measurement Errors)

	ρ Value				
	-0.9	-0.5	0	0.5	0.9
			Coefficient on Main Variable		
PCA	0.316	0.778	1.0	1.223	1.684
	(0.027)	(0.023)	(0.021)	(0.024)	(0.028)
Single Measurement	0.243	0.714	1.0	1.286	1.757
	(0.025)	(0.024)	(0.022)	(0.024)	(0.026)
All Measurements	0.316	0.778	1.0	1.223	1.684
	(0.027)	(0.023)	(0.021)	(0.024)	(0.028)
Average of Measurements	0.316	0.778	1.0	1.223	1.684
	(0.027)	(0.023)	(0.021)	(0.024)	(0.028)
Instrumental Variable	0.348	0.801	1.0	1.202	1.652
	(0.05)	(0.033)	(0.029)	(0.032)	(0.043)
			Mean Absolute Percentage Error		
PCA	68.4%	22.2%	1.7%	22.3%	68.4%
Single Measurement	75.7%	28.6%	1.7%	28.6%	75.7%
All Measurements	68.4%	22.2%	1.7%	22.3%	68.4%
Average of Measurements	68.4%	22.2%	1.7%	22.3%	68.4%
Instrumental Variable	65.2%	19.9%	2.2%	20.2%	65.2%
Simulations	1,000	1,000	1,000	1,000	1,000

We can see that all methods that use multiple covariate measurements perform worse than in the uncorrelated measurement error case, with IV having an especially dramatic change. PCR still performs better than using a single measurement, and in line with using all measurements or the average of the measurements, while now performing only slightly worse than IV. In fact, the average coefficient for PCR is within a single standard deviation of the IV coefficient for every ρ value studied. Additionally, IV has larger standard errors than PCR (and the other techniques) under this measurement error framerwork. As can be seen in the simulation appendix, these findings roughly hold for other values of β_1, β_2, and ρ (although larger p values seem to erode any advantage that IV holds over PCR).

Overall, these simulations tell us that PCR helps to reduce measurement error-induced bias in OLS regression with a latent covariate relative to including only a single measurement of that covariate. Additionally, it seems to perform in line with taking the average of those covariate measurements and using each measurement as a separate covariate in the regression. However, PCR performs much worse than an instrumental variable regression when the covariate measurement errors are uncorrelated with one another, and ever-so-slightly worse when covariate measurement errors are correlated.

Application: Government Share of Healthcare Spending and Life Expectancy

We now examine the usage of the principal components estimator in an empirical setting with measurement error. One interesting question in public economics and public health is the study of the relationship between publicly and privately funded healthcare systems and outcomes such as life expectancy. To measure the public or private nature of a healthcare system we use the continuous variable of the government's share of total health expenditure in a given country and year.

Some previous work has covered this relationship. Considering that this topic has been studied in "relatively few papers," Linden and Ray (2017) focus on the relationship between life expectancy at birth and public and private health expenditures for 34 OECD countries from 1970-2012 and find that both public and private health spending are important to life expectancy and are associated with each other. In work similar to ours, Or (2000) predicts premature death in 21 OECD countries from 1970-1992, considering the public share of health expenditure, environmental factors, and GDP. He finds that a larger share of public spending is associated with lower rates of premature mortality for both males and females, and that controlling for GDP is important; it is also associated with less premature mortality. This work also demonstrates the importance of our methods of reducing the number of covariates considered, as it includes many economic variables and fixed effects but examines only several hundred observations; the estimators used may be subject to an significant amount of variance.

In this regression it is important to account for the role of a country's level of economic development. There is an extensive literature documenting the relationship between economic development and life expectancy. Ling et al. (2017) finds that economic growth is associated with increased life expectancy in Malaysia, while considering the reverse causal direction Acemoglu and Johnson (2007) finds improvements in life expectancy lead to little or no growth. Somewhat less obvious is the linkage between government provision of healthcare and development. In general, public goods provision and government spending, including in fields such as healthcare, has been linked to prosperity; low income countries may remain in such a state due to inefficient governments and inferior institutions (Wu, Tang and Lin, 2010).

However, economic development is liable to be measured with error. GDP measurements usually rely on company surveys, and the methodology within a country and for comparisons between countries through exchange rates or PPP adjustments may vary (Grishin, Ustyuzhanina and Pavlovna, 2019). Other sources of error include the presence of the informal economy and non-monetary but productive work, the challenge of accurately measuring the value of digital services which often do not have visible prices, and government incentives to manipulate official statistics about growth (Charmes, 2012; Ahmad, Ribarsky and Reinsdorf, 2017; Nakamura, Steinsson and Liu, 2016).[1] Hence, this setup, with a covariate in regression subject to measurement error, fits the situation described in the theory and simulations in the previous sections.[2] In this case, we aim to to reduce possible bias in the coefficient of the government's share of health spending by making appropriate use of multiple measures of economic development.

Our data on all measures comes from the World Bank (The World Bank, 2021). We remove country-years with missing values for any of the variables summarized in Table 4 and standardize the economic covariates in the 5 uppermost rows by subtracting the mean and dividing by the standard deviation (in order to enable interpretable principal component analysis).

[1]Due to differences in statistical capacity and the larger relative size of the informal economy, it is possible that mismeasurement of economic development is particularly severe in developing countries. On the other hand, the presence of the digital economy may mean mismeasurement is larger in developed nations. This would constitute the presence of non-classical measurement error, but we only consider classical measurement error in this paper. It is also possible that the interaction of many forms of measurement produces error which is closer to classical assumptions.

[2]It seems less likely that our variable of interest, the government's share of health spending, is measured with error. One would think most governments capable of monitoring their own spending better than economic activity in general. Furthermore, for this variable there would seem to be less governmental incentive to manipulate the statistic relative to GDP or other items.

Table 4: Summary Statistics

Variable	Obs	Mean	SD	Min	Med	Max
GDP Per Capita PPP (Current International $)	3,143	16,443.76	19,173.04	435.08	9,331.99	141,634.96
GDP Per Capita (Current USD)	3,143	11,764	17,582.7	111.93	4,018.95	118,823.65
GNP Per Capita PPP (Current International $)	3,143	15,980.12	18,478.14	410	9,080	132,440
GNP Per Capita (Current USD)	3,143	11,247.65	16,583.07	110	3,800	104,560
ILO GDP Per Person Employed	3,143	41,782.32	39,743.08	1,371.24	29,220.02	266,103.71
Life Expectancy at Birth (All Population)	3,143	69.55	9.26	39.44	71.78	84.21
Government Share of Health Expenditure	3,143	49.62	21.73	4.06	50.27	95.14

In Table 5 we apply our main set of estimators. The PCA estimator is found in column (1). In column (2), we instead control for just a single measurement of GDP per capita (PPP). In column (3) we directly include the full set of economic covariates (measurements) listed in the 5 uppermost rows of Table 4. In column (4), we use the mean of these mismeasured covariates, and in column (5) we perform instrumental variables regression using all the other economic indicators as an instrument for GDP per capita (PPP).

Table 5: Regressions of Life Expectancy on Government Share of Health Spending

	Life Expectancy at Birth (Years)				
	(1)	(2)	(3)	(4)	(5)
Govt. Share of Health Exp.	0.180^{***} (0.007)	0.194^{***} (0.007)	0.166^{***} (0.007)	0.180^{***} (0.007)	0.193^{***} (0.007)
Covariates	PCA	Single Measurement (GDP Per Capita PPP)	All Measurements	Average of Measurements	Instrumental Variable (GDP Per Capita PPP)
Observations	3,143	3,143	3,143	3,143	3,143
Note:					*p<0.1; **p<0.05; ***p<0.01

The PCA estimator indicates that a one percentage point increase in the government share of health expenditure is linked to a increase in life expectancy of 0.18 years. Notably, the results generally demonstrate that the methods using multiple measures of the covariates produce different coefficients for government health share, relative to the use of a single mismeasurement. The principal components estimator, the usage of many economic indicators directly, and their mean each produce far smaller coefficients. The IV estimator produces a slightly smaller coefficient.

Moreover, these different coefficients behave in a manner similar to that predicted by our theoretical development and simulations. In Table 2, we saw the impact of variation in ρ, the correlation between measurements for a values of $p = 5$ and $\beta_1, \beta_2 = 1$ for 3,000 observations. In the empirical setting it is difficult to tell what is a reasonable value of β. Nevertheless, we see that for a positive ρ value between 0 and 1 (as is likely to be the case in light the correlation between GDP and the government share of health spending and overall public goods), the coefficient obtained from using a single measurement is inflated relative to that from PCA, and presumably other methods combining multiple measures as in columns (3), (4), and (5) of Table 5.

Results using univariate OLS (with no covariates), country and year fixed effects models (with country clustered standard errors), and more principal components are in Table 6. Univariate OLS produces a large and inflated coefficient. Fixed effects coefficients greatly reduce the magnitude of any potential causal effects and are insignificant. The results in column (4) also show a reduction in the inflation of coefficients relative to column (2) of Table 5, as the inclusion of more principal components produces a small coefficient very similar to that obtained with just a single component. [3]

[3] This is likely due to the high explanatory power of just the first principal component, as is clear in Appendix Figure 3.

Table 6: Additional Regressions

	Life Expectancy at Birth (Years)			
	(1)	(2)	(3)	(4)
Govt. Share of Health Exp.	0.276***	-0.003	-0.003	0.181***
	(0.006)	(0.011)	(0.011)	(0.007)
Covariates	None	None	PCA	PC 1-2
Fixed Effects	No	Yes	Yes	No
Observations	3,143	3,143	3,143	3,143
Note:		*p<0.1; **p<0.05; ***p<0.01		

Conclusion

In this paper, we have shown the usefulness of using Principal Component Regression (PCR) as a method of mitigating the bias on the coefficient for the main variable of interest that is induced by measurement error in an auxiliary covariate. Using theory and Monte Carlo simulations we have shown how this technique improves upon using a single measurement of the true covariate, and we have shown how its performance compares to that of some other methods for mitigating this bias, while mostly falling short of the performance of IV. We then applied this technique to an empirical question (the relationship between the government share of health expenditure and life expectancy). Our findings suggest that PCR is a useful tool for dealing with issues caused by measurement error in an auxiliary covariate.

Other solutions to measurement error meriting further exploration include the usage of general factor models, which could also involve the exploitation of the panel structure of relevant data. Measurements within time periods or units may provide information about the true value of variables. There are also likely other ways to summarize and control the dimension of the information presented by multiple measurements. Another option could be the usage of dimensionality-reduction techniques within an instrumental variables regression.

References

Acemoglu, Daron, and Simon Johnson. 2007. "Disease and Development: The Effect of Life Expectancy on Economic Growth." *Journal of Political Economy*, 115(6): 925–985. Publisher: The University of Chicago Press.

Ahmad, Nadim, Jennifer Ribarsky, and Marshall Reinsdorf. 2017. "Can potential mismeasurement of the digital economy explain the post-crisis slowdown in GDP and productivity growth?" Publisher: OECD.

Charmes, Jacques. 2012. "The Informal Economy Worldwide: Trends and Characteristics." *Margin: The Journal of Applied Economic Research*, 6(2): 103–132. Publisher: SAGE Publications India.

Grishin, Victor Ivanovich, Elena Vladimirovna Ustyuzhanina, and Irina Pavlovna. 2019. "Main Problems with Calculating GDP as and Indicator of Economic Health of the County." 9.

Heckman, James, Susanne Schennach, and Benjamin D Williams. 2010. "Matching on Proxy Variables." 15.

Linden, Mikael, and Deb Ray. 2017. "Life expectancy effects of public and private health expenditures in OECD countries 1970–2012: Panel time series approach." *Economic Analysis and Policy*, 56: 101–113.

Ling, Chong Hui, Khalid Ahmed, Rusnah Muhamad, Muhammad Shahbaz, and Nanthakumar Loganathan. 2017. "Testing the Social Cost of Rapid Economic Development in Malaysia: The Effect of Trade on Life Expectancy." *Social Indicators Research*, 130(3): 1005–1023.

Nagasawa, Kenichi. 2020. "Identification and Estimation of Partial Effects with Proxy Variables." 23.

Nakamura, Emi, Jón Steinsson, and Miao Liu. 2016. "Are Chinese Growth and Inflation Too Smooth? Evidence from Engel Curves." *American Economic Journal: Macroeconomics*, 8(3): 113–144.

Or, Zeynep. 2000. "Determinants of Health Outcomes in Industrialised Countries: a Pooled, Cross-country, Time-series Analysis." *OECD Economic Studies*, 25.

Schennach, Susanne M. 2016. "Recent Advances in the Measurement Error Literature." *Annual Review of Economics*, 8(1): 341–377.

Schofield, Lynne Steuerle. 2015. "Correcting for Measurement Error in Latent Variables Used as Predictors." *The annals of applied statistics*, 9(4): 2133–2152.

The World Bank. 2021. "Indicators | Data."

Wegge, Leon L. 1996. "Local identifiability of the factor analysis and measurement error model parameter." *Journal of Econometrics*, 70(2): 351–382.

Wu, Shih-Ying, Jenn-Hong Tang, and Eric S. Lin. 2010. "The impact of government expenditure on economic growth: How sensitive to the level of development?" *Journal of Policy Modeling*, 32(6): 804–817.

Theory Appendix

Short Proofs

Claim 1 *Under measurement error in the covariates, the coefficient is such that:*

$$\gamma^* = \gamma + \beta \frac{\text{cov}(t,x)(\sigma^2_{x^*} - \sigma^2_x)}{\sigma^2_t \sigma^2_{x^*} - \text{cov}(t,x)^2} \tag{15}$$

Proof of claim 1.1 (1) *Then, equations* (4) *and* (5) *will be such that*

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} \sigma^2_t & \text{cov}(t,x^*) \\ \text{cov}(x^*,t) & \sigma^2_{x^*} \end{pmatrix}^{-1} \begin{pmatrix} \text{cov}\,(y,t) \\ \text{cov}\,(y,x^*) \end{pmatrix} \tag{16}$$

$$\tag{17}$$

Claim 2 *Suppose* $Z_i = X_i + \omega_i$. *If* $\mathrm{E}(\omega_i) = 0$, $\mathrm{Cov}(\epsilon_i, \omega_i) = 0$ *and* $\mathrm{Cov}(\eta_i, \omega_i) = 0$. *Then*

$$\mathrm{E}(Z_i \zeta_i) = 0 \text{ and } \mathrm{E}(Z_i X_i) \neq 0 \tag{18}$$

And so γ *can be identified through IV regression.*

Proof of claim 2.1 *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*

$$Z_i' y_i = Z_i' \gamma^* t_i + Z_i' X_i^{*\prime} \beta^* + Z_i' \zeta_i \tag{19}$$

and so

$$\begin{pmatrix} \gamma^{IV} \\ \beta^{IV} \end{pmatrix} = \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} + \Sigma_{Z\eta} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{20}$$

$$= \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{21}$$

$$\begin{pmatrix} \gamma^{IV} \\ \beta^{IV} \end{pmatrix} = \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{22}$$

However, finding a reliable source of exogeneity is sometimes difficult, as is demonstrating the exclusion restriction. Suppose now that as instrument we have another measure of X_i *so that*

$$Z_i = X_i + \omega_i \tag{23}$$

where $\mathrm{E}(\omega_i) = 0$, $\mathrm{Cov}(\epsilon_i, \omega_i) = 0$ *and that* ω_i *brings new information so that* $\mathrm{Cov}(\eta_i, \omega_i) = 0$. *Then, if* Z_i *satisfies exogeneity and relevance we will be able to identify the parameters without any bias as shown in equations* (13) *to* (15). *In* *fact:*

$$\mathrm{E}(Z_i\zeta_i) = \mathrm{E}(Z_i(\epsilon_i - \eta_i\beta) \tag{24}$$
$$= \mathrm{E}(Z_i\epsilon_i) - \mathrm{E}(Z_i\eta_i)\beta \tag{25}$$
$$= \mathrm{E}((X_i + \omega_i)\epsilon_i) - \mathrm{E}((X_i + \omega_i)\eta_i)\beta \tag{26}$$
$$= \mathrm{E}(X_i\epsilon_i) + \mathrm{E}(\omega_i\epsilon_i) - (\mathrm{E}(X_i\eta_i) + \mathrm{E}(\omega_i\eta_i))\beta \tag{27}$$

And so Z_i is exogenous. Given (16) it is clear that $\mathrm{E}(Z_iX_i) \neq 0$ and so relevance is also satisfied. Thus, γ and β may be identifies using this kind of instrument.

Claim 3

$$X_i = HX_i^* \tag{28}$$

where H is a matrix compound of the h_k vectors of eigenvalues of $x_{i,k}$, $\forall i, k$

Proof of claim 3.1 *Under our assumptions, the vector of mismeasured 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*

$$x_{i,k} = h_k' X_k^* \tag{29}$$

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

$$X_i = HX_i^* \tag{30}$$

where H is a matrix such that

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

and h_k is the vector of eigenvalues for the variable $x_{i,k}$.

Proof of claim 3.2 *See Theory Appendix*

Claim 4 *Consider equation (11). Then*

$$\gamma^{PCR} = \gamma \tag{31}$$

Proof of claim 4.1 *The coefficients are as follows*

$$\begin{pmatrix} \gamma^{PCR} \\ \beta^{PCR} \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{y,HX^*} \end{pmatrix} \tag{32}$$

$$= \begin{pmatrix} \sigma_t^2 & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_{X} \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{33}$$

$$= \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{34}$$

where the last equality comes from (13)

Simulation Appendix

Table 7: Average Coefficients for Values of β_1 (Uncorrelated Measurement Errors)

	True β_1		
	0.1	1	10
	Coefficient on Main Variable		
PCA	0.205	1.105	10.106
	(0.021)	(0.023)	(0.022)
Single Measurement	0.385	1.286	10.287
	(0.023)	(0.024)	(0.023)
All Measurements	0.205	1.105	10.106
	(0.021)	(0.023)	(0.022)
Average of Measurements	0.205	1.105	10.106
	(0.021)	(0.023)	(0.022)
Instrumental Variable	0.1	1.0	10.0
	(0.044)	(0.038)	(0.043)
	Mean Absolute Percentage Error		
PCA	105.1%	10.5%	1.1%
Single Measurement	285.3%	28.6%	2.9%
All Measurements	105.1%	10.5%	1.1%
Average of Measurements	105.0%	10.5%	1.1%
Instrumental Variable	32.1%	3.0%	0.3%
Simulations	1,000	1,000	1,000

Table 8: Average Coefficients for Values of β_2 (Uncorrelated Measurement Errors)

	True β_2		
	0.1	1	10
	Coefficient on Main Variable		
PCA	1.012	1.105	2.055
	(0.021)	(0.023)	(0.081)
Single Measurement	1.03	1.286	3.864
	(0.02)	(0.024)	(0.127)
All Measurements	1.012	1.105	2.055
	(0.021)	(0.023)	(0.081)
Average of Measurements	1.012	1.105	2.055
	(0.021)	(0.023)	(0.081)
Instrumental Variable	1.001	1.0	1.009
	(0.029)	(0.038)	(0.3)
	Mean Absolute Percentage Error		
PCA	1.9%	10.5%	105.5%
Single Measurement	3.1%	28.6%	286.4%
All Measurements	1.9%	10.5%	105.5%
Average of Measurements	1.9%	10.5%	105.5%
Instrumental Variable	2.2%	3.0%	22.6%
Simulations	1,000	1,000	1,000

Table 9: Average Coefficients for Values of p (No Transformation of Measurements)

	Number of p		
	5	20	50
	Coefficient on Main Variable		
PCA	1.105	1.032	1.013
	(0.023)	(0.021)	(0.022)
Single Measurement	1.286	1.286	1.286
	(0.024)	(0.023)	(0.024)
All Measurements	1.105	1.032	1.013
	(0.023)	(0.021)	(0.022)
Average of Measurements	1.105	1.032	1.013
	(0.023)	(0.021)	(0.022)
Instrumental Variable	1.0	1.005	1.01
	(0.038)	(0.031)	(0.031)
	Mean Absolute Percentage Error		
PCA	10.5%	3.3%	2.0%
Single Measurement	28.6%	28.6%	28.6%
All Measurements	10.5%	3.3%	2.0%
Average of Measurements	10.5%	3.3%	2.0%
Instrumental Variable	3.0%	2.5%	2.6%
Simulations	1,000	1,000	1,000

Table 10: Average Coefficients for Values of β_1 (Correlated Measurement Errors)

	True β_1		
	0.1	1	10
	Coefficient on Main Variable		
PCA	0.323	1.223	10.222
	(0.022)	(0.024)	(0.023)
Single Measurement	0.387	1.286	10.286
	(0.022)	(0.024)	(0.023)
All Measurements	0.323	1.223	10.222
	(0.022)	(0.024)	(0.023)
Average of Measurements	0.323	1.223	10.222
	(0.022)	(0.024)	(0.023)
Instrumental Variable	0.3	1.202	10.2
	(0.034)	(0.032)	(0.038)
	Mean Absolute Percentage Error		
PCA	223.0%	22.3%	2.2%
Single Measurement	286.7%	28.6%	2.9%
All Measurements	223.0%	22.3%	2.2%
Average of Measurements	223.0%	22.3%	2.2%
Instrumental Variable	200.8%	20.2%	2.0%
Simulations	1,000	1,000	1,000

Table 11: Average Coefficients for Values of β_2 (Correlated Measurement Errors)

	True β_2		
	0.1	1	10
	Coefficient on Main Variable		
PCA	1.022	1.223	3.222
	(0.02)	(0.024)	(0.118)
Single Measurement	1.028	1.286	3.856
	(0.02)	(0.024)	(0.13)
All Measurements	1.022	1.223	3.222
	(0.02)	(0.024)	(0.118)
Average of Measurements	1.022	1.223	3.222
	(0.02)	(0.024)	(0.118)
Instrumental Variable	1.019	1.202	3.002
	(0.03)	(0.032)	(0.199)
	Mean Absolute Percentage Error		
PCA	2.5%	22.3%	222.2%
Single Measurement	3.0%	28.6%	285.6%
All Measurements	2.5%	22.3%	222.2%
Average of Measurements	2.5%	22.3%	222.2%
Instrumental Variable	2.7%	20.2%	200.2%
Simulations	1,000	1,000	1,000

Table 12: Average Coefficients for Values of p (Correlated Measurement Errors)

	Number of p		
	5	20	50
	Coefficient on Main Variable		
PCA	1.223	1.206	1.202
	(0.024)	(0.023)	(0.023)
Single Measurement	1.286	1.286	1.285
	(0.024)	(0.023)	(0.023)
All Measurements	1.223	1.206	1.202
	(0.024)	(0.023)	(0.023)
Average of Measurements	1.223	1.206	1.202
	(0.024)	(0.023)	(0.023)
Instrumental Variable	1.202	1.201	1.201
	(0.032)	(0.025)	(0.024)
	Mean Absolute Percentage Error		
PCA	22.3%	20.6%	20.2%
Single Measurement	28.6%	28.6%	28.5%
All Measurements	22.3%	20.6%	20.2%
Average of Measurements	22.3%	20.6%	20.2%
Instrumental Variable	20.2%	20.1%	20.1%
Simulations	1,000	1,000	1,000

Application Appendix

Figure 1: Correlations Between Variables

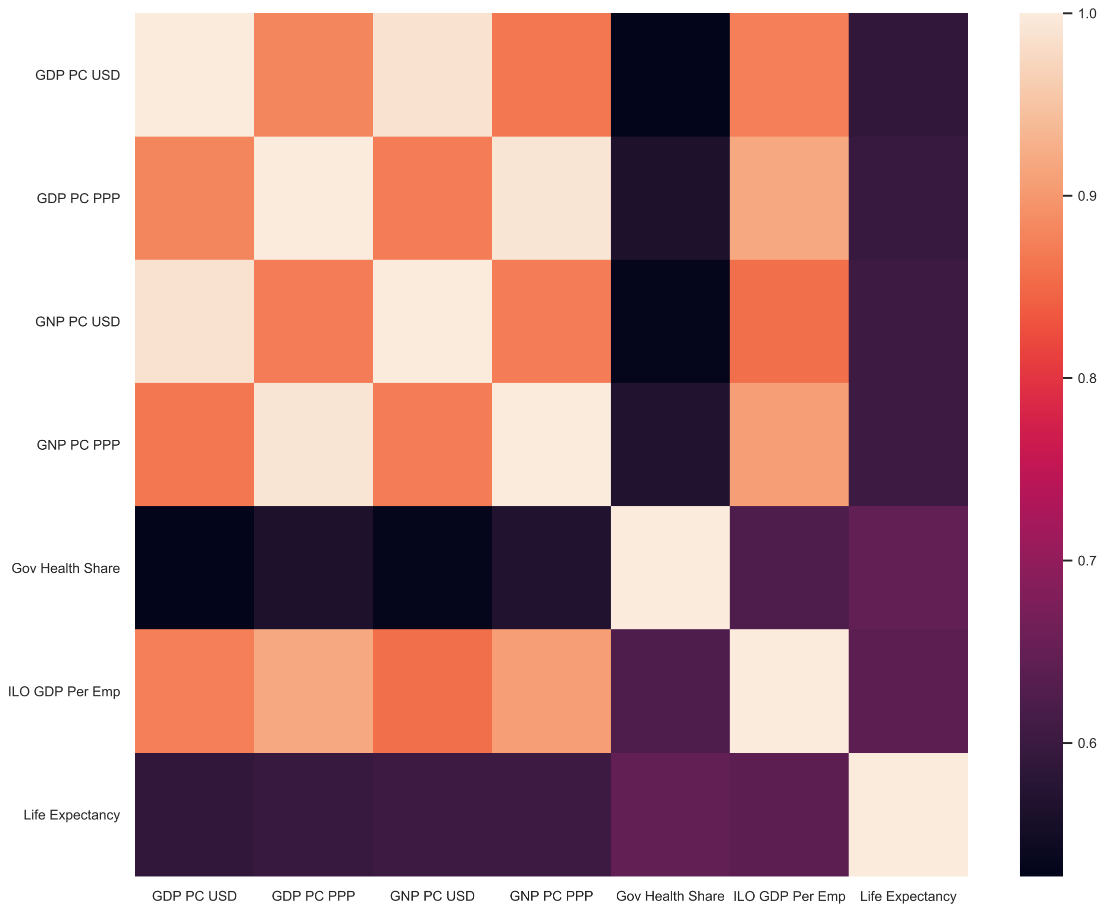

Figure 2: Economic Measures PCA Loadings

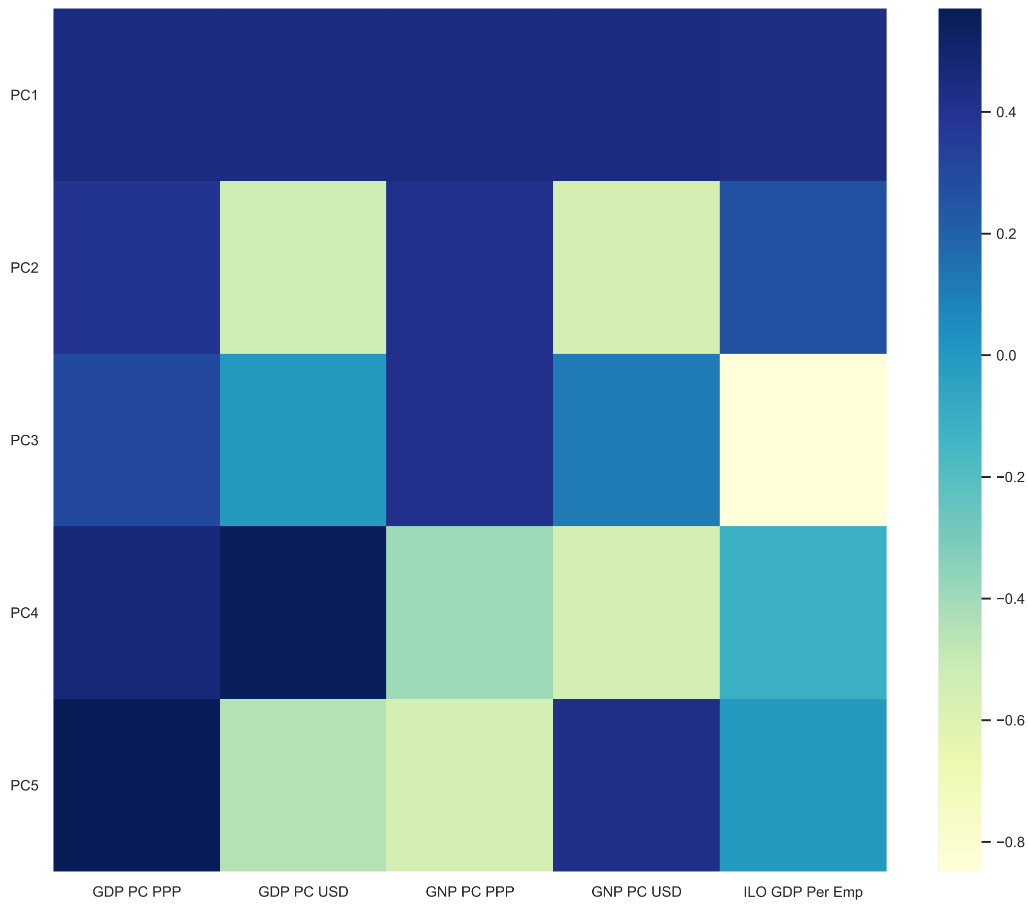

Figure 3: Economic Measures PCA Share of Variance Explained