Principal Component Regression as a Solution to Measurement Error Bias

Isaac Liu, Nicolás Martorell & Paul Opheim

May 29, 2021

Abstract

We argue that using Principal Component Regression (PCR) is a useful solution to bias introduced by measurement error in the covariate 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 then apply this method 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, 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. One may then use this extracted value in a standard OLS regression, thus providing a solution to attenuation bias that does not require the strong 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. In this case, these 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, in order 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 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, Household Income Per Capita, 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 missmeasured 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 $X_i = (x_{i,1}, x_{i,2}, \ldots x_{i,p})'$. 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 $V(x_{i,k}^*) = \sigma_{x_k}^2 + \sigma_{\eta_k}^2 \geq \sigma_{x_k}^2$.

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 $\epsilon_i \sim \text{iid}(0, \sigma_\epsilon^2)$. 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

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

the coefficients would be such that

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

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 the coefficient of our variable of interest will be biased.

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

$$\gamma^* = \gamma + \beta \frac{\text{cov}(t, x)(\sigma_{x^*}^2 - \sigma_x^2)}{\sigma_t^2 \sigma_{x^*}^2 - \text{cov}(t, x)^2} \tag{6}$$

Proof of claim 1.1 *See Appendix 4*

From (10) it is clear that when $\text{cov}(t, x) \neq 0$ and that x is measured with error (i.e $\sigma_{x^*}^2 > \sigma_x^2$), 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 $\sigma_{x^*}^2 = \sigma_x^2$ and so we would not be facing 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, $\sigma_{x^*}^2 > \sigma_x^2$ which implies $\sigma_{x^*}^2 - \sigma_x^2 > 0$. 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 one illustrates.

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

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

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 $\text{E}(Z_i'X_i) \neq 0$ and $\text{E}(Z_i't_i) \neq 0$, and also the exclusion restriction $\text{E}(Z_i'\epsilon_i) = \text{E}(Z_i'\zeta_i) = \text{E}(Z_i'\eta_{i,k}) = 0$, for all i and k. Then premultiplying by Z_i we have

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

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.

Suppose now that as instrument we have another measure of X_i so that

$$Z_i = X_i + \omega_i \tag{11}$$

where $\text{E}(\omega_i) = 0$, $\text{Cov}(\epsilon_i, \omega_i) = 0$ and that ω_i brings new information so that $\text{Cov}(\eta_i, \omega_i) = 0$. 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:

$$\text{E}(Z_i\zeta_i) = \text{E}(Z_i(\epsilon_i - \eta_i\beta) \tag{12}$$

$$= \text{E}(Z_i\epsilon_i) - \text{E}(Z_i\eta_i)\beta \tag{13}$$

$$= \text{E}((X_i + \omega_i)\epsilon_i) - \text{E}((X_i + \omega_i)\eta_i)\beta \tag{14}$$

$$= \text{E}(X_i\epsilon_i) + \text{E}(\omega_i\epsilon_i) - (\text{E}(X_i\eta_i) + \text{E}(\omega_i\eta_i))\beta \tag{15}$$

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

Alternatively, we propose an alternative bias-correction method when there are several missmeasured 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

$$h_j = \underset{h'h=1,h'h_1=0,\ldots,h'h_{j-1}=0}{\mathrm{argmax}} \mathrm{var}\left[h'X_k^*\right] \tag{16}$$

where h_j is the eigenvector of Σ associated with the jth 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 jth 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

$$x_{i,k} = h_k'X_k^* \tag{17}$$

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

$$X_i = HX_i^* \tag{18}$$

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

$$y_i = \gamma^{PCR}t_i + HX_i^{*\prime}\beta^{PCR} + \epsilon_i \tag{19}$$

where the coefficients are as follows

$$\begin{pmatrix} \gamma^{PCR} \\ \beta^{PCR} \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{y,HX^*} \end{pmatrix} \tag{20}$$

$$= \begin{pmatrix} \sigma_t^2 & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{21}$$

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

where the last equality comes from (13).

Note that according to equation (17), the true variable $x_{i,k}$ is a linear combination of the missmeasured variables that the researcher may have, were the weights are such that equation (16) is satisfied. This allows us to think about other linear combination that could be used as a bias-correction method. In particular, take the case in which $h_k =' \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$, where h_k is a row vector of dimension $(1 \times J)$, and J is the amount of missmeasured

variables for $x_{i,k}$. Then, equation (17) will be

$$\tilde{x_{i,k}} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} x_{i,1}^* \\ x_{i,2}^* \\ \vdots \\ x_{i,J}^* \end{pmatrix} \tag{23}$$

$$= \frac{1}{n} \sum_{j=1}^{J} x_{i,j}^* \tag{24}$$

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

Estimation of the Principal Component Regression

Recall that the missmeasured variables are such that $x_{i,k}^* = x_{i,k} + \eta_{i,k}$. Then for every i, X_i^* can we get

$$X_i^* = X_i + \mu \tag{25}$$

and so we could interpret the vector of missmeasured variables X_i^* as a factor model, in which the common factor is the vector of true variables X_i. This way, we could rewrite the model as follows

$$y_i = \gamma t_i + X_i' \beta + \epsilon_i \tag{26}$$
$$X_i^* = X_i + \mu \tag{27}$$

Which is a factor-augmented regression model in which the common "factors" between the missmeasured variables of $x_{i,k}$ is in fact $x_{i,k}$. We estimate the model following two-stages. First, we estimate X_i by factor regression. Then, the first-stage is the principal-components estimation

$$\hat{X}_i = \hat{D}^{-1} X_i^* \tag{28}$$

Second step is to regress Y_i on the estimated $\hat{X}_i$. Then, the coefficients would be

$$\begin{pmatrix} \hat{\gamma}^F \\ \hat{\beta}^F \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_t^2 & \hat{\Sigma}_{t,\hat{X}} \\ \hat{\Sigma}_{\hat{X},t} & \hat{\Sigma}_{\hat{X}} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\Sigma}_{yt} \\ \hat{\Sigma}_{y,\hat{X}} \end{pmatrix} \tag{29}$$

Using the *Frisch-Waugh-Lowell* decomposition

$$\hat{\gamma}^F = ((M_{\hat{X}} t)' M_{\hat{X}} t)^{-1} (M_{\hat{X}} t)' M_{\hat{X}} y \tag{30}$$

Where $M_{\hat{X}} = I - \hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}'$. Then

$$\hat{\gamma}^F = ((M_{\hat{X}} t)' M_{\hat{X}} t)^{-1} (M_{\hat{X}} t)' M_{\hat{X}} (t\gamma + X\beta + \epsilon) \tag{31}$$
$$= \gamma + ((M_{\hat{X}} t)' M_{\hat{X}} t)^{-1} (M_{\hat{X}} t)' M_{\hat{X}} X\beta + ((M_{\hat{X}} t)' M_{\hat{X}} t)^{-1} (M_{\hat{X}} t)' M_{\hat{X}} \epsilon \tag{32}$$

Properties of the Estimator: Monte Carlo Simulations

We then 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 $\mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$, where ρ is some 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 many 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 Σ is a diagonal p by p matrix with only 1s on the diagonal.

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 2,000 observations of y_i, x_i, etc. Table 2 shows the results for different values of p.

Table 2: Average Coefficients for Values of p (No Transformation of Measurements)

	5	20	50
	Number of p		
	Coefficient on Main Variable		
PCA	1.104	1.030	1.014
	(0.027)	(0.026)	(0.026)
Single Measurement	1.284	1.285	1.287
	(0.029)	(0.029)	(0.028)
All Measurements	1.104	1.030	1.013
	(0.027)	(0.022)	(0.018)
Average of Measurements	1.104	1.030	1.013
	(0.027)	(0.022)	(0.017)
Instrumental Variable	1.125	1.037	1.029
	(0.028)	(0.026)	(0.026)
	Absolute Percentage Error		
PCA	10.4%	3.3%	2.3%
	(2.7 ppts)	(2.2 ppts)	(1.8 ppts)
Single Measurement	28.4%	28.5%	28.7%
	(2.8 ppts)	(2.9 ppts)	(2.9 ppts)
All Measurements	10.4%	3.3%	2.3%
	(2.7 ppts)	(2.2 ppts)	(1.8 ppts)
Average of Measurements	10.4%	3.3%	2.3%
	(2.7 ppts)	(2.2 ppts)	(1.7 ppts)
Instrumental Variable	12.5%	3.9%	3.2%
	(2.8 ppts)	(2.4 ppts)	(2.2 ppts)
Observations	1,000	1,000	1,000

We can see that using PCA to extract the latent covariate driving the mismeasured covariates noticeably outperforms using a single mismeasured covariate across several values of p. Both the average coefficient on β_1 obtained when including the PCA output in the regression, and the mean absolute percentage error obtained on the 1,000 simulations are both much closer to the target values with the PCA-based regression than with the single measurement regression. Additionally, we can see that as p increases the estimated β_1 coefficient in the PCA regression gets steadily closer to the true β_1 value of 1. Appendix 1 contains charts that show that this increase in

performance is also true for different values of β_1 and β_2.

However, the performance advantages that we see from using PCA could be driven by the benefit of having multiple measurements of our true covariate of interest, as opposed to any special advantages from PCA specifically. We test this question by comparing the estimated β_1 in our PCA 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 p are also shown in Table 2.

As one can see from these results (and results for different values of β_1, β_2, and ρ in Appendix 1), there does not seem to be a noticeable difference between the first three regression methods, with the instrumental variable regression performing slightly worse. Thus, our simulations suggest that there are major benefits to having multiple measurements of a latent covariate of interest, but under our first framework, using PCA, taking the average of these measurements, and including all measurements as separate covariates seem to give similar benefits to the performance of the regression.

Additionally, there are certain circumstances where the PCA method does not lead to more accurate estimates of β_1 than even using a single mismeasured covariate. Table 3 contains the simulation results for different values of ρ (the covariance between the main variable of interest x_i and the true latent covariate z_i).

Table 3: Average Coefficients for Values of ρ (No Transformation of Measurements)

	ρ Value				
	-0.9	-0.5	0	0.5	0.9
	Coefficient on Main Variable				
PCA	0.538	0.894	1.000	1.104	1.464
	(0.039)	(0.027)	(0.024)	(0.027)	(0.041)
Single Measurement	0.244	0.714	1.000	1.284	1.759
	(0.030)	(0.029)	(0.027)	(0.028)	(0.031)
	Absolute Percentage Error				
PCA	46.2%	10.6%	1.9%	10.4%	46.4%
	(3.9 ppts)	(2.7 ppts)	(1.4 ppts)	(2.7 ppts)	(4.1 ppts)
Single Measurement	75.6%	28.6%	2.2%	28.4%	75.9%
	(3.0 ppts)	(2.9 ppts)	(1.6 ppts)	(2.8 ppts)	(3.1 ppts)
Observations	1,000	1,000	1,000	1,000	1,000

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 1.28 instead of the true value 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 is hardly any 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.

We also explore how our results differ under a different framework of measurement error. For the following regressions (and those in Appendix 2) we transform half of the covariate measurements by taking the $z_{i,j}^*$ value from before and transforming it so that the new $z_{i,j}'$ value was equal to $z_{i,j}' = e^{z_{i,j}^*}$. This allows us to analyze the performance of our different techniques in a situation where the measurements of the true covariate are on

different scales from one another. [1] Table 4 shows the results of our five techniques in this new context.

Table 4: Average Coefficients for Values of p (Half of Measurements are Transformed)

	Number of p		
	5	20	50
Coefficient on Main Variable			
PCA	1.208	1.082	1.054
	(0.032)	(0.029)	(0.028)
Single Measurement	1.286	1.286	1.286
	(0.029)	(0.029)	(0.028)
All Measurements	1.182	1.057	1.024
	(0.029)	(0.027)	(0.026)
Average of Measurements	1.365	1.303	1.285
	(0.034)	(0.029)	(0.028)
Instrumental Variable	1.234	1.065	1.039
	(0.029)	(0.027)	(0.026)
Absolute Percentage Error			
PCA	20.8%	8.2%	5.5%
	(3.2 ppts)	(2.9 ppts)	(2.7 ppts)
Single Measurement	28.6%	28.6%	28.6%
	(2.9 ppts)	(2.9 ppts)	(2.8 ppts)
All Measurements	18.2%	5.8%	2.9%
	(2.9 ppts)	(2.6 ppts)	(2.1 ppts)
Average of Measurements	36.5%	30.3%	28.5%
	(3.4 ppts)	(3.3 ppts)	(3.2 ppts)
Instrumental Variable	23.4%	6.5%	4.1%
	(2.9 ppts)	(2.7 ppts)	(2.4 ppts)
Observations	1,000	1,000	1,000

We can see that taking the average of the covariate measurements now performs noticeably worse than PCA or using all measurements as separate covariates. Additionally, IV techniques seem to perform better in this paradigm and attains results consistent with those for the PCA and all measurements techniques. As can be seen in Appendix 2, these findings hold for other values of β_1, β_2, and ρ.

Thus, these Monte Carlo simulations tell us that using PCR provides significant performance benefits relative to using a single mismeasured covariate of the true latent covariate in a regression. When the measurements of the true covariate have the same scale and only have Gaussian measurement error, then PCR performs about as well as taking the average of all covariate measurements or as using all measurements as separate covariates in an OLS regression. When we took the exponential of half of the measurements, then PCR outperformed taking the average of the covariate measurements and performed in line with using all measurements as covariates or using an IV technique. These simulations thus tell us that PCR is a useful way of reducing measurement error-caused bias in OLS regression with a latent covariate, and that the performance of this technique is in line with, or better than, other techniques for reducing that bias.

Application: Government Share of Healthcare Spending and Life Expectancy

We now examine the implications 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.

---

[1]This is often the case; for example, one could measure income inequality through a Gini coefficient and the percent of income that goes to the top 1% of income-earners, but these variables are on totally different scales from each other.

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).[2] 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.[3] 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 standardize all variables by subtracting the mean and dividing by the standard deviation, and remove country-years with missing values for any of the variables summarized in Table 6.

In Table 5 below we begin with a univariate OLS regression, which produces a large and significant coefficient indicating a one standard deviation increase in the government share of health expenditure is linked to a 0.647 standard deviation increase in life expectancy. Next we include the potentially mismeasured covariate of GDP per capita, which greatly reduces the size of the coefficient on the governments share. After this we include a full set of economic covariates listed in the 5 uppermost rows of Table 6, again reducing the size of the coefficient. However, this result is then adjusted upwards again in the last two columns, where use the mean of the mismeasured covariates, and the first principal component.

---

[2]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.

[3]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

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.	0.647***	0.455***	0.388***	0.421***	0.421***
	(0.014)	(0.015)	(0.016)	(0.015)	(0.015)
Covariates	None	GDP PC PPP	Econ Indicators	Mean	PC 1
Observations	3,143	3,143	3,143	3,143	3,143
R^2	0.418	0.497	0.528	0.515	0.515
Adjusted R^2	0.418	0.497	0.527	0.515	0.515
Residual Std. Error	0.763	0.710	0.688	0.697	0.697
F Statistic	2257.875***	1551.203***	584.979***	1669.297***	1668.832***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6: Summary Statistics

Variable	Obs	Mean	SD	Min	Med	Max
GDP Per Capita PPP (Current International $)	3143	16463.45	19208.93	435.08	9331.99	141634.96
GDP Per Capita (Current USD)	3143	11764.00	17582.70	111.93	4018.95	118823.65
GNP Per Capita PPP (Current International $)	3143	15999.02	18509.76	410.00	9080.00	132440.00
GNP Per Capita (Current USD)	3143	11247.65	16583.07	110.00	3800.00	104560.00
ILO GDP Per Person Employed	3143	41852.00	39834.76	1371.24	29254.99	266953.37
Life Expectancy at Birth (All Population)	3143	69.55	9.26	39.44	71.78	84.21
Government Share of Health Expenditure	3143	49.62	21.73	4.06	50.27	95.14

The results clearly demonstrate that all of the methods using multiple measures of the covariates produce noticably different coefficients for government health share. The usage of many economic indicators directly, the mean, and the principal components estimator each produce smaller coefficients.

Moreover, these different coefficients behave in a manner similar to that predicted by our theoretical development and simulations. In Table 3, we saw the impact of variation in ρ, the correlation between measurements for a value of $p = 5$ and $\beta = 1$ for 100 observations. In the empirical setting it is difficult to tell what is a reasonable value of β, and we instead have nearly 2000 observations. 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 fixed effects panel models (with country clustered standard errors), more principal components, and instrumental variables regression using all the economic indicators as an instrument for GDP per capita, are in Table 7. Fixed effects coefficients greatly reduce the magnitude of any potential causal effects and are insignificant. The results in column (3) and (4) also show reductions in the inflation of coefficients relative to column (2) of Table 5. The inclusion of more principal components produces a small coefficient similar to that obtained with just a single component, while though the reduction provided by the IV coefficient is minimal.

seem to be less governmental incentive to manipulate the statistic relative to GDP or other items.

Table 7: Additional Regressions

	Life Expectancy at Birth (Years)			
	(1)	(2)	(3)	(4)
Govt. Share of Health Exp.	-0.006	-0.008	0.424***	0.453***
	(0.026)	(0.026)	(0.015)	(0.015)
Covariates	None	PC 1	PC 1-2	GDP PC (IV)
Fixed Effects	Yes	Yes	No	No
Observations	3,143	3,143	3,143	3,143
R^2	0.980	0.979	0.516	0.497
Adjusted R^2	0.978	0.978	0.515	0.497
Residual Std. Error	0.147	0.148	0.696	0.709
F Statistic	10250.216***	22274.480***	1114.058***	nan***

Note: *p<0.1; **p<0.05; ***p<0.01

Conclusion

In this paper, we have

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.

Appendix 1

Table 8: Average Coefficients for Values of β_1 (No Transformation of Measurements)

	True β_1			
	0.1	1	10	
	Coefficient on Main Variable			
PCA	0.205	1.104	10.106	
	(0.026)	(0.027)	(0.027)	
Single Measurement	0.384	1.284	10.286	
	(0.028)	(0.028)	(0.029)	
All Measurements	0.205	1.104	10.106	
	(0.026)	(0.027)	(0.027)	
Average of Measurements	0.205	1.104	10.106	
	(0.026)	(0.027)	(0.027)	
Instrumental Variable	0.225	1.125	10.126	
	(0.027)	(0.028)	(0.028)	
	Absolute Percentage Error			
PCA	104.9%	10.4%	1.1%	
	(26.1 ppts)	(2.7 ppts)	(0.3 ppts)	
Single Measurement	284.4%	28.4%	2.9%	
	(28.2 ppts)	(2.8 ppts)	(0.3 ppts)	
All Measurements	104.7%	10.4%	1.1%	
	(26.1 ppts)	(2.7 ppts)	(0.3 ppts)	
Average of Measurements	104.7%	10.4%	1.1%	
	(26.1 ppts)	(2.7 ppts)	(0.3 ppts)	
Instrumental Variable	125.4%	12.5%	1.3%	
	(26.6 ppts)	(2.8 ppts)	(0.3 ppts)	
Observations	1,000	1,000	1,000	1,000

Table 9: Average Coefficients for Values of β_2 (No Transformation of Measurements)

	True β_1			
	0.1	1	10	
	Coefficient on Main Variable			
PCA	1.011	1.104	2.057	
	(0.025)	(0.027)	(0.106)	
Single Measurement	1.029	1.284	3.863	
	(0.024)	(0.028)	(0.159)	
All Measurements	1.011	1.104	2.054	
	(0.025)	(0.027)	(0.106)	
Average of Measurements	1.011	1.104	2.054	
	(0.025)	(0.027)	(0.106)	
Instrumental Variable	1.013	1.125	2.259	
	(0.025)	(0.028)	(0.116)	
	Absolute Percentage Error			
PCA	2.2%	10.4%	105.7%	
	(1.6 ppts)	(2.7 ppts)	(10.6 ppts)	
Single Measurement	3.2%	28.4%	286.3%	
	(2.0 ppts)	(2.8 ppts)	(15.9 ppts)	
All Measurements	2.2%	10.4%	105.4%	
	(1.6 ppts)	(2.7 ppts)	(10.6 ppts)	
Average of Measurements	2.2%	10.4%	105.4%	
	(1.6 ppts)	(2.7 ppts)	(10.6 ppts)	
Instrumental Variable	2.3%	12.5%	125.9%	
	(1.6 ppts)	(2.8 ppts)	(11.6 ppts)	
Observations	1,000	1,000	1,000	1,000

Table 10: Average Coefficients for Values of ρ (No Transformation of Measurements)

	ρ Value				
	-1	-0.5	0	0.5	1
Coefficient on Main Variable					
PCA	0.538	0.894	1.000	1.104	1.464
	(0.039)	(0.027)	(0.024)	(0.027)	(0.041)
Single Measurement	0.244	0.714	1.000	1.284	1.759
	(0.030)	(0.029)	(0.027)	(0.028)	(0.031)
All Measurements	0.539	0.894	1.000	1.104	1.463
	(0.039)	(0.027)	(0.014)	(0.027)	(0.041)
Average of Measurements	0.539	0.894	1.000	1.104	1.463
	(0.039)	(0.027)	(0.014)	(0.027)	(0.041)
Instrumental Variable	0.487	0.874	1.000	1.125	1.514
	(0.038)	(0.027)	(0.024)	(0.028)	(0.040)
Absolute Percentage Error					
PCA	46.2%	10.6%	1.9%	10.4%	46.4%
	(3.9 ppts)	(2.7 ppts)	(1.4 ppts)	(2.7 ppts)	(4.1 ppts)
Single Measurement	75.6%	28.6%	2.2%	28.4%	75.9%
	(3.0 ppts)	(2.9 ppts)	(1.6 ppts)	(2.8 ppts)	(3.1 ppts)
All Measurements	46.1%	10.6%	1.9%	10.4%	46.3%
	(3.9 ppts)	(2.7 ppts)	(1.4 ppts)	(2.7 ppts)	(4.1 ppts)
Average of Measurements	46.1%	10.6%	1.9%	10.4%	46.3%
	(3.9 ppts)	(2.7 ppts)	(1.4 ppts)	(2.7 ppts)	(4.1 ppts)
Instrumental Variable	51.3%	12.6%	2.0%	12.5%	51.4%
	(3.8 ppts)	(2.7 ppts)	(1.4 ppts)	(2.8 ppts)	(4.0 ppts)
Observations	1,000	1,000	1,000	1,000	1,000

Table 11: Average Coefficients for Values of β_1 (Half of Measurements are Transformed)

	True β_1		
	0.1	1	10
Coefficient on Main Variable			
PCA	0.307	1.208	10.205
	(0.032)	(0.032)	(0.032)
Single Measurement	0.385	1.286	10.284
	(0.029)	(0.029)	(0.028)
All Measurements	0.281	1.181	10.180
	(0.028)	(0.029)	(0.027)
Average of Measurements	0.465	1.365	10.363
	(0.034)	(0.034)	(0.033)
Instrumental Variable	0.333	1.234	10.232
	(0.029)	(0.029)	(0.028)
Absolute Percentage Error			
PCA	207.3%	20.8%	2.1%
	(32.1 ppts)	(3.2 ppts)	(0.3 ppts)
Single Measurement	285.0%	28.6%	2.8%
	(28.6 ppts)	(2.9 ppts)	(0.3 ppts)
All Measurements	181.3%	18.2%	1.8%
	(27.9 ppts)	(2.9 ppts)	(0.3 ppts)
Average of Measurements	364.6%	36.5%	3.6%
	(34.2 ppts)	(3.4 ppts)	(0.3 ppts)
Instrumental Variable	233.3%	23.4%	2.3%
	(28.8 ppts)	(2.9 ppts)	(0.3 ppts)
Observations	1,000	1,000	1,000

Table 12: Average Coefficients for Values of β_2 (Half of Measurements are Transformed)

	True β_1			
	0.1	1	10	
	Coefficient on Main Variable			
PCA	1.021	1.208	3.091	
	(0.025)	(0.032)	(0.214)	
Single Measurement	1.029	1.286	3.871	
	(0.024)	(0.029)	(0.158)	
All Measurements	1.018	1.182	2.826	
	(0.025)	(0.029)	(0.129)	
Average of Measurements	1.036	1.365	4.669	
	(0.024)	(0.034)	(0.247)	
Instrumental Variable	1.023	1.234	3.347	
	(0.024)	(0.029)	(0.160)	
	Absolute Percentage Error			
PCA	2.7%	20.8%	209.1%	
	(1.8 ppts)	(3.2 ppts)	(21.4 ppts)	
Single Measurement	3.2%	28.6%	287.1%	
	(2.0 ppts)	(2.9 ppts)	(15.8 ppts)	
All Measurements	2.5%	18.2%	182.6%	
	(1.8 ppts)	(2.9 ppts)	(12.9 ppts)	
Average of Measurements	3.8%	36.5%	366.9%	
	(2.1 ppts)	(3.4 ppts)	(24.7 ppts)	
Instrumental Variable	2.8%	23.4%	234.7%	
	(1.9 ppts)	(2.9 ppts)	(16.0 ppts)	
Observations	1,000	1,000	1,000	1,000

Table 13: Average Coefficients for Values of ρ (Half of Measurements are Transformed)

	ρ Value				
	-1	-0.5	0	0.5	1
	Coefficient on Main Variable				
PCA	0.337	0.793	0.997	1.208	1.663
	(0.041)	(0.033)	(0.026)	(0.032)	(0.041)
Single Measurement	0.244	0.716	0.997	1.286	1.757
	(0.031)	(0.029)	(0.028)	(0.029)	(0.031)
All Measurements	0.377	0.819	0.997	1.182	1.624
	(0.037)	(0.028)	(0.026)	(0.029)	(0.036)
Average of Measurements	0.177	0.637	0.998	1.365	1.822
	(0.030)	(0.034)	(0.029)	(0.034)	(0.030)
Instrumental Variable	0.302	0.768	0.998	1.234	1.698
	(0.035)	(0.030)	(0.027)	(0.029)	(0.035)
	Absolute Percentage Error				
PCA	66.3%	20.7%	2.1%	20.8%	66.3%
	(4.1 ppts)	(3.3 ppts)	(1.6 ppts)	(3.2 ppts)	(4.1 ppts)
Single Measurement	75.6%	28.4%	2.3%	28.6%	75.7%
	(3.1 ppts)	(2.9 ppts)	(1.7 ppts)	(2.9 ppts)	(3.1 ppts)
All Measurements	62.3%	18.1%	2.1%	18.2%	62.4%
	(3.6 ppts)	(2.8 ppts)	(1.6 ppts)	(2.9 ppts)	(3.6 ppts)
Average of Measurements	82.2%	36.3%	2.3%	36.5%	82.2%
	(3.0 ppts)	(3.4 ppts)	(1.8 ppts)	(3.4 ppts)	(3.0 ppts)
Instrumental Variable	69.8%	23.2%	2.1%	23.4%	69.8%
	(3.5 ppts)	(3.0 ppts)	(1.6 ppts)	(2.9 ppts)	(3.5 ppts)
Observations	1,000	1,000	1,000	1,000	1,000

Appendix 2

Appendix 3

Figure 1: Correlations Between Covariates and Life Expectancy
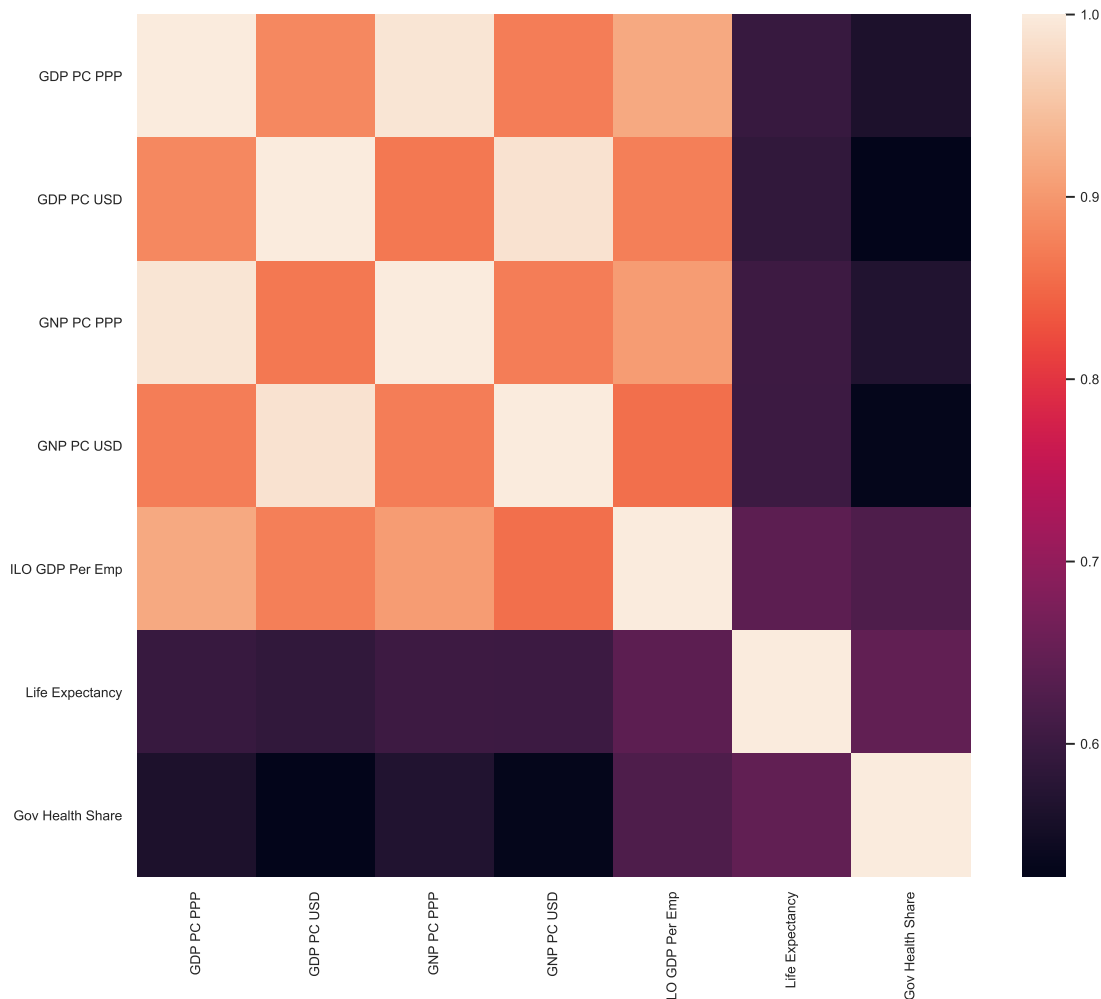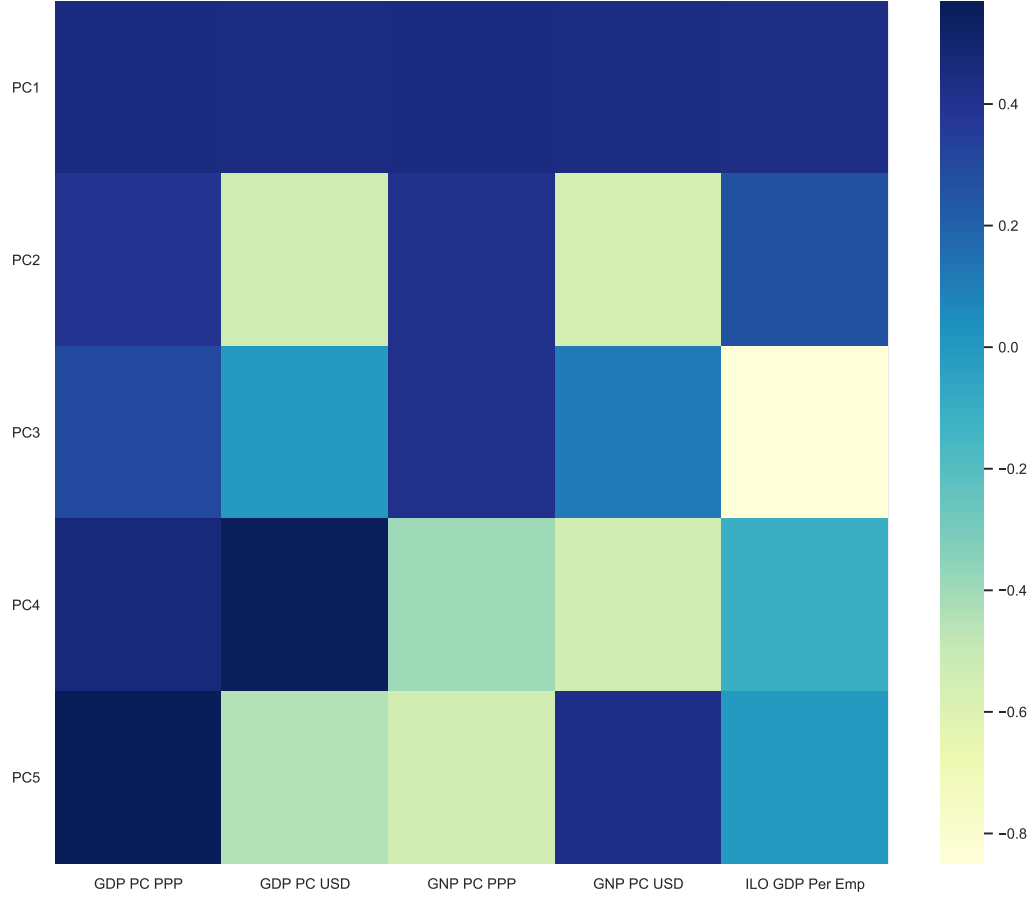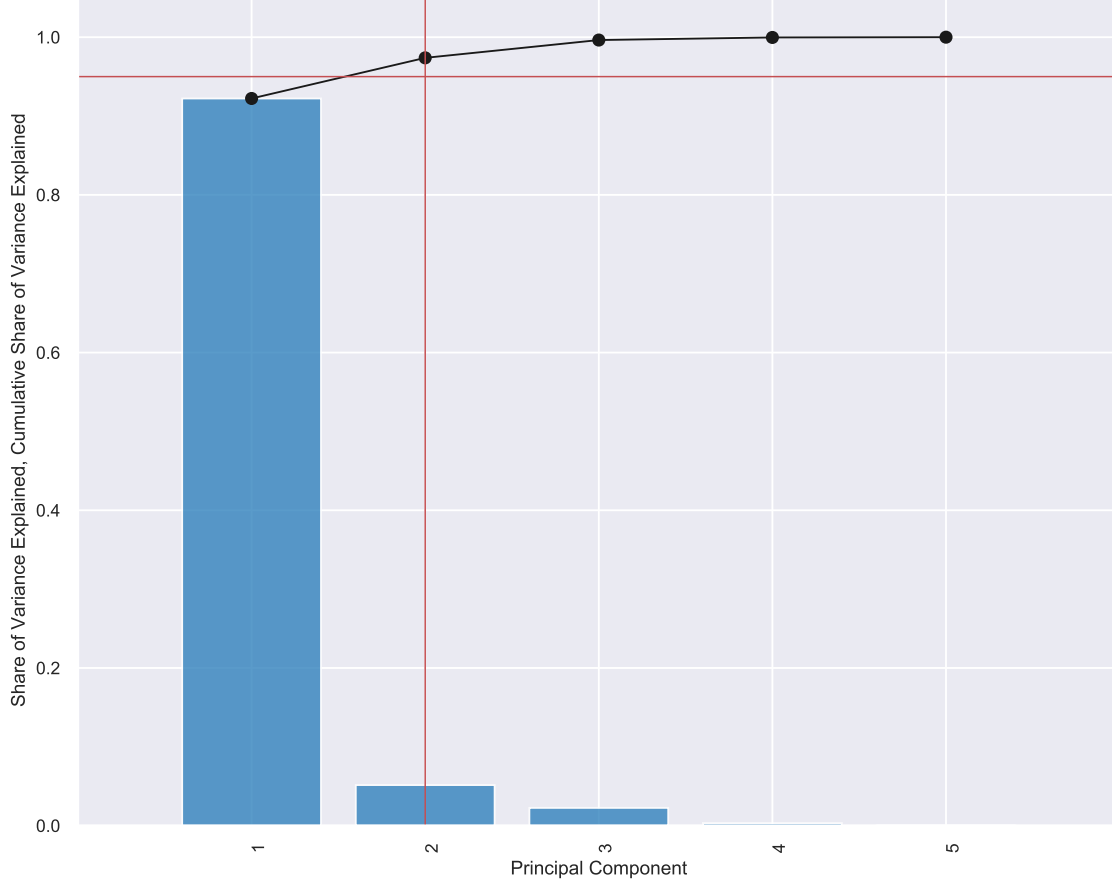
Figure 2: Economic Measures PCA Loadings

Appendix 4

Short Proofs

Claim 2 *Under measurement error in the covariates, the coefficient is such that:*

$$\gamma^* = \gamma + \beta \frac{\operatorname{cov}(t,x)(\sigma_{x^*}^2 - \sigma_x^2)}{\sigma_t^2 \sigma_{x^*}^2 - \operatorname{cov}(t,x)^2} \tag{33}$$

Figure 3: Economic Measures PCA Share of Variance Explained

Proof of claim 2.1 *Then, equations (4) and (5) will be such that*

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \mathrm{cov}(t, x^*) \\ \mathrm{cov}(x^*, t) & \sigma_{x^*}^2 \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{cov}(y, t) \\ \mathrm{cov}(y, x^*) \end{pmatrix} \tag{34}$$

$$\tag{35}$$

Thus, the coefficient of our variable of interest will be

$$\gamma^* = \frac{\sigma_{x^*}^2 \, \mathrm{cov}(y, t) - \mathrm{cov}(t, x^*) \, \mathrm{cov}(y, x^*)}{\sigma_t^2 \sigma_{x^*}^2 - \mathrm{cov}(t, x^*)^2} \tag{36}$$