Identification of a Triangular Two Equation System Without Instruments

Arthur Lewbel, Susanne M. Schennach, and Linqi Zhang*
Boston College, Brown University, and Boston College

Original September 2020, Revised December 2022

Abstract

We show that a standard linear triangular two equation system can be point identified, without the use of instruments or any other side information. We find that the only case where the model is not point identified is when a latent variable that causes endogeneity is normally distributed. In this non-identified case, we derive the sharp identified set. We apply our results to Acemoglu and Johnson's (2007) model of life expectancy and GDP, obtaining point identification and comparable estimates to theirs, without using their (or any other) instrument.

*__Keywords__: Returns to schooling, identification, triangular Systems, Kotlarski, deconvolution. Corresponding Author: Arthur Lewbel, Department of Economics - Maloney 315, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://sites.google.com/bc.edu/arthur-lewbel/

1 Introduction

Consider a standard linear triangular structural model

$$Y = X'b_1 + \varepsilon_1 \tag{1}$$

$$W = \gamma Y + X'b_2 + \varepsilon_2 \tag{2}$$

for some endogenous variables Y and W, exogenous covariates X, and unobserved errors ε_1 and ε_2. For example, W could be a worker's wages or earnings and Y could be her level of schooling. Or, as in our later empirical application, W could be a country's GDP growth and Y a health measure like growth in life expectancy. The primary goal is identification of γ, the direct causal effect of Y on W, though we will also obtain identification of b_1, b_2, and the joint distribution of the errors.[1]

The main obstacle to identification and estimation of γ is that ε_1 and ε_2 may be correlated, because both depend on a common unobserved U (ability in the case of schooling and wages, technology in the case of GDP and health). That is, in its simplest form,

$$\varepsilon_1 = U + V \quad \text{and} \quad \varepsilon_2 = \beta U + R \tag{3}$$

where U, V, and R are unobserved, mutually independent (conditional on X) random variables and β is a constant. After projecting off covariates X, the V and R errors represent idiosyncratic shocks to Y and W, while U is what makes Y an endogenous regressor in the W equation.

Similar triangular structural models arise whenever we have one variable Y affecting another variable W, and a common unobservable that affects them both. For example, consider a two period dynamic model with autocorrelated errors. In this case W equals Y in a subsequent time period, and U represents the autocorrelation in the errors. Another example is production, where W could be a firm's value-added output per unit of capital, Y

[1] Throughout this paper we focus on the traditional homogeneous effects model where γ is a constant, rather than a heterogeneous treatment effects model.

is the firm's labor per unit of capital, and U is unobserved entrepreneurship, which affects both productivity and the chosen level of inputs.

Such models are traditionally identified in econometrics by finding an instrument, i.e., a variable that correlates with Y but not ε_2, or equivalently, a variable that correlates with V but not U or R. However, such instruments can be difficult to find. For example, Card (1995, 2002) and others propose using measures of access to schooling, such as distance to or cost of colleges in one's area, as wage equation instruments, while others raise objections to the validity of these instruments, e.g. Carneiro and Heckman (2002). Other wage equation instruments may raise fewer questions of validity but can be weak, like Angrist and Krueger's (1991, 2001) quarter of birth instruments.

Similarly, Acemoglu and Johnson (2007) propose using changes in predicted mortality, constructed based on innovations in health care, as an instrument for life expectancy growth Y in their regression of GDP growth W on Y. However, such health innovations could be correlated with other technological advances that increase GDP, leading to instrument invalidity. Comparable questions can be raised regarding the instruments or identifying side information in other similar studies, such as Aghion, Howitt, and Murtin (2010), who find a positive γ, in contrast to Acemoglu and Johnson's (2007) negative γ. Ecevit (2013) summarizes results from eleven similar studies, finding estimates of γ that range from strongly negative to insignificant to strongly positive. This range of estimates raises serious questions regarding the validity of instruments or other side information that different authors use to identify γ.[2]

Rather than propose any new instrument, we address the more fundamental question of whether and when this model can be point identified and estimated *without* side information such as instruments whose validity can be hard to ascertain (noting that the alternative of a randomized experiment is not feasible for a macro question like this). If so, then we

[2]Of course, differences are also due to variation in data sets and in how Y and W are defined and constructed. As another way to explain these differing results, Cervellati and Sunde (2011) suggest that the true effect might be non-monotonic.

can estimate the model without relying on side information, and/or test the validity of side information like instruments via overidentification tests.

We provide conditions for point identification of the model

$$Y = U + V \tag{4}$$

$$W = \gamma Y + \beta U + R \tag{5}$$

with U, V, and R being unobserved, mutually independent random variables with unknown distributions. The same identification theorem can then be applied conditioning on covariates X, to show point identification of more general models, where the entire distributions of U, V, and R could depend nonparametrically on X. A special case of this general identification result is then identification of equations (1), (2) and (3). In this special case, variables V and R that depend nonparametrically on X in equations (4) and (5) are instead replaced with $X'\beta_1 + V$ and $X'\beta_2 + R$, where these new V and R do not depend on X.[3]

Our main result is surprising: If the sign of β is known a priori, under minimal regularity assumptions, the coefficients γ and β, and the distributions of U, V, and R (and b_1 and b_2 in that model) are all point identified without instruments or other side information, unless either U or V is normally distributed (after appropriately conditioning on or projecting off covariates X). So, for example, Y having bounded support would be a sufficient condition for point identification, since that would rule out normality of U or V.

In addition to proving this general identification result, we also: 1. Provide a few low order moments yielding simple GMM estimators of the model, 2. Show how infinitely many additional moments conditions can be systematically constructed to provide identification under weaker conditions, 3. Provide the sharp identified set for the coefficients γ and β in the case where either U or V is normal and hence point identification fails, 4. Investigate the behavior of these GMM estimators in some Monte Carlo exercises, and 5. Provide

[3]More generally, U, V and R could be heteroskedastic, or otherwise have higher moments that depend in unknown ways on X, but estimation would then become more complicated. One possibility would be replacing the GMM estimators we provide with conditional moment GMM, conditioning on X. More simply, heteroskedasticity could be parameterized, with parameters estimated as part of the GMM.

an empirical application where we establish that our identification and estimation strategy is viable even with a very small sample size. Specifically, we estimate the Acemoglu and Johnson (2007) model without using any instruments, and obtain estimates that are very similar to what they found with their instrument.

Instrumental variables estimation of the model has the advantage that it only requires assumptions regarding first and second moments of the covariates, errors, and instruments. In contrast, our assumptions regarding U, V, and R are, implicitly, restrictions on all moments. However, there are a number of mitigating factors. First, some of our results, such as Lemma 1 below, only rely on lower order moments. Second, our main theorem works via convolutions, and so our independence assumptions can be relaxed to subindependence, as defined and described in Schennach (2019), who points out that subindependence is arguably as weak as a conditional mean assumption in terms of the dimensionality of the restrictions imposed. Third, our independence assumption is actually conditional on other covariates, so, e.g., the identification can handle arbitrary heteroskedasticity and dependence of higher moments on regressors. Similarly, if, e.g., U is ability, then identification only requires ability to be conditionally (sub)independent from other unobserved factors, conditional on covariates. Nevertheless, given our required assumptions, these results should be most useful when instruments either don't exist, or might be invalid.

The identification of equations (4) and (5) without instruments has been previously considered by Rigobon (2003), Klein and Vella (2010), and Lewbel (2012), but these results neither nest nor are nested by ours because they *require* that the errors be heteroskedastic, and identification is obtained by imposing varying restrictions on the structure of that heteroskedasticity.[4]

A number of special cases of our results do appear in the literature, but all of them assume $\gamma = 0$, and so they omit the most important feature of the model in applications like ours.

[4]Rigobon (2003) and Klein and Vella (2010) impose different parametric restrictions on the error variances, while Lewbel (2012) imposes a nonparametric restriction. For simplicity we assume homoskedastic errors, but by conditioning our identification theorems on X, we could allow for general heteroskedastity as well, at the expense of likely weaker identification and more complicated estimators.

Kotlarski (1967) is the special case of our model where it is known that $\gamma = 0$ and $\beta = 1$, and in that case Kotlarski's Lemma shows that point identification of the distribution of all the latent variables holds even under normality. Similarly, Reiersøl (1950) uses a special case of our model where it is known that $\gamma = 0$ and Y plays the role of a measurement of U contaminated by an error V and establishes conditions under which β would be identified. As noted in Lewbel (2020), with $\gamma = 0$ and Reiersøl's identification of β, one could rewrite Reiersøl's model as $Y = U + V$ and $W/\beta = U + R/\beta$, and then apply Kotlarski's lemma to the joint distribution of Y and W/β to identify the distributions of U, V, and R.[5]

Our results, showing necessary and sufficient conditions to identify the more general model of equations (4) and (5) with unknown nonzero γ, turns out to be a difficult extension. In particular, the methods of proof used by Reiersøl (1950) and Kotlarski (1967) do not extend to our problem. The proof of our main result instead relies on similar tools as Khatri and Rao (1972) or Rao (1966, 1971) (see also Comon's (1994) reference to Darmois (1953)).

Some limitations of our results should be acknowledged upfront. We assume that the coefficients γ and β are constants. So, e.g., our results do not immediately extend to random coefficients, such as treatment effects with unobserved heterogeneity, or to nonlinearity in the dependence of W on Y. However, this limitation may be mitigated to some extent by allowing the distributions of the unobservables to be unknown functions of covariates. Another important restriction on our results is that we require U to be a scalar. While this is a common assumption (as in the examples cited earlier), there are other situations where one might expect a vector of unobservable shocks like U to affect both Y and W, and our identification results would then not apply. We provide examples in Supplement D. Finally,

[5]A special case of non-normality is when the components U and V are asymmetric. Lewbel (1997) and Erickson and Whited (2002) exploit asymmetry to construct simple estimators for the Reiersøl (1950) model. See also Bierens (1981). Other papers propose estimators for models like equations (4) and (5) with $\gamma = 0$, by assuming that coefficients like β are point identified using higher moments, but without explicitly characterizing when that is possible. Examples include Bonhomme and Robin (2010), Fruehwirth, Navarro, and Takahashi (2016), and Navarro and Zhou (2017). A related result, showing identification of direction of causality in models under nonnormality, is Peters, Janzing, and Scholkopf (2017). Generalizations of Kotlarski's lemma to models with more components (but again still assuming $\gamma = 0$) include Székely and Rao (2000) and Li and Zheng (2020). A nonlinear extension of Reiersøl (1950) is Schennach and Hu (2013).

a limitation for empirical work is that our estimators depend on higher than second moments of the data, and such moments can lead to very imprecise estimates when sample sizes are small.

In section 2, we provide a few simple moments that will often suffice to point identify our model, and can be used to construct a correspondingly simple GMM estimator. In Section 3, we present our general identification results, including constructing more moments like those in Section 2, and showing that, with minimal regularity, the model is point identified as long as both U and V are not normal. In sections 4 and 5 we derive the sharp identified set when either U or V is normal, and derive some inequalities regarding our model relative to ordinary least squares. Section 6 provides a Monte Carlo analysis of our simple GMM estimators. In section 7 we provide an empirical application based on Acemoglu and Johnson (2007), in which we obtain estimates comparable to theirs, without using their (or any other) instrument. Section 8 concludes with some suggestions for further work.

2 Simple Identification and Estimation

We begin with a simple special case of our general results, by providing some moments that can easily be used to identify and estimate (by standard GMM) the models described in the introduction. These results are not as general as our main identification theorem, but are likely to suffice for many empirical applications.

We first consider identification and estimation of equations (4) and (5) without covariates X, and then we extend the results to equations (1) and (2).

Assumption 1 *We observe the joint distribution of two real valued, nondegenerate random variables Y and W.*

With data, we could assume independent, identically distributed observations of Y and W, and then identify their joint distribution to satisfy Assumption 1 using the Glivenko Cantelli theorem.

Assumption 2 *The unobserved real valued random variables U, V, and R are mean zero and mutually independent,*[6] *with unknown distributions.*

Assumption 3 *R has finite variance, and U and V each have finite fourth moments.*

Assumption 4 *The unknown constants γ and β are real valued, finite, and $\beta > 0$.*

We can assume our data Y and W have been demeaned, rationalizing the assumption that the unobservables have mean zero. To see why we need a sign restriction on β, observe that we can rearrange equations (4) and (5) to get $W = (\gamma + \beta) Y - \beta V + R$, which, except for the sign of β, is observationally equivalent to the original model, switching the roles of V and U. Usually, the sign of β should be clear from the economics of the application, e.g., in a returns to schooling model, $\beta > 0$ is a natural assumption, since it says that unobserved ability that increases (decreases) education outcomes will increase (decrease) wages. If we instead believed β was negative, we could just replace Y with $-Y$ everywhere to make β positive (redefining γ, U, and V accordingly).

We also rule out $\beta = 0$, because if $\beta = 0$ then it would be pointless to separately identify V and U. Moreover, having $\beta = 0$ is nonsensical in the types of applications we consider, since it would mean that Y is exogenous, making identification and estimation of γ trivial.

Substituting equation (4) into equation (5) gives the reduced form expression for W

$$W = \gamma V + \alpha U + R \quad \text{with} \quad \alpha = \gamma + \beta \tag{6}$$

The following Lemma provides two moments that can often suffice to point identify γ and α, which then trivially also point identifies β.

Lemma 1 *Let Assumptions 1-4 and equations (4) and (5) (and therefore also equation 6) hold. Then*

$$E\left[\left(W - \gamma Y\right)\left(W - \alpha Y\right) Y\right] = 0 \tag{7}$$

$$cov\left[\left(W - \gamma Y\right)\left(W - \alpha Y\right), Y^2\right] - 2E\left(WY - \gamma Y^2\right) E\left(WY - \alpha Y^2\right) = 0 \tag{8}$$

[6] Independence can be weakened to subindependence (Schennach (2019)).

Proofs are all in Supplement A. The proof of Lemma 1 works by substituting $W - \gamma Y = \beta U + R$ and $W - \alpha Y = -\beta V + R$ into equations (7) and (8), and then uses the mutual independence of U, V, and R to verify that these equations hold.

Lemma 1 provides two equations in the two unknowns α and γ. If we solve the first equation for α and substitute that into the second, we obtain a quadratic in γ. The sign restriction that $\beta > 0$ then determines which root is the correct one for γ.

We later provide the formal conditions under which these two equations suffice to point identify α and γ. The main condition, derived in Theorem 1 below, is equation (21). Equation (21) shows that the main cases in which equations (7) and (8) by themselves fail to provide point identification are when U and V have the exact same distribution, or when both are symmetrically distributed, or if either U or V is normally distributed. We later show that infinitely many additional equations in α, γ, Y and W can be constructed, based on higher moments of Y and W than those used in Lemma 1. These higher moments can help identify α and γ in applications where Lemma 1 does not suffice.

A simple estimator for α and β can be constructed by rewriting equations (7) and (8) as moment conditions, and applying standard method of moments or GMM. One can immediately check that these equations take the form

$$E\left(YW - \mu_{yw}\right) = 0, \qquad E(Y^2 - \mu_{yy}) = 0 \tag{9}$$

$$E\left[\left(W - \gamma Y\right)\left(W - \left(\gamma + \beta\right)Y\right)Y\right] = 0 \tag{10}$$

$$E\left[\left(W - \gamma Y\right)\left(W - \left(\gamma + \beta\right)Y\right)\left(Y^2 - \mu_{yy}\right) - 2\left(\mu_{yw} - \gamma\mu_{yy}\right)\left(W - \left(\gamma + \beta\right)Y\right)Y\right] = 0 \tag{11}$$

where $\mu_{yw} = E\left(YW\right)$ and $\mu_{yy} = E\left(Y^2\right)$. The parameters μ_{yw} and μ_{yy} are estimated along with γ and β by putting equations (9), (10), and (11) into any standard GMM estimation routine. One could replace β with e^b in these equations to impose the sign restriction that $\beta > 0$.

Lemma 1 uses up to fourth moments of the data. Based on results derived in the next section, in Supplement B we provide additional equations (using up to fifth moments) that

can provide overidentification of γ and β, or point identification in some cases where Lemma 1 does not suffice.

Let σ_U^2, σ_V^2, and σ_R^2 denote the variances of the error components U, V, and R. It may be of economic interest to estimate these variances, to identify how much of the variance of the model errors is due to unobserved ability U versus the idiosyncratic components V and R. From the model we have $E\left(\left(W-\gamma Y\right)Y\right) = \beta\sigma_U^2$, $E\left(Y^2\right) = \sigma_U^2 + \sigma_V^2$, and $E\left(\left(W-\gamma Y\right)^2\right) = \beta^2\sigma_U^2 + \sigma_R^2$, which implies

$$\sigma_U^2 = E\left(\left(W-\gamma Y\right)Y\right)/\beta, \qquad \sigma_V^2 = E\left(Y^2\right) - \sigma_U^2, \qquad \sigma_R^2 = E\left(\left(W-\gamma Y\right)^2\right) - \beta^2\sigma_U^2 \tag{12}$$

Given estimates of β and γ, we can replace the expectations in equation (12) with sample averages to estimate these variances.

Alternatively, we can estimate these variances jointly with the model parameters by observing that

$$\mu_{yy} = \sigma_U^2 + \sigma_V^2, \qquad \mu_{yw} = \beta\sigma_U^2 + \gamma\left(\sigma_U^2 + \sigma_V^2\right). \tag{13}$$

So, in equations (9), (10), and (11) we can replace μ_{yy} and μ_{yw} with their expressions in equation (13), and apply GMM using those equations along with the additional equation

$$E\left(\left(W-\gamma Y\right)^2 - \beta^2\sigma_U^2 - \sigma_R^2\right) = 0 \tag{14}$$

to simultaneously estimate β, γ, σ_U^2, σ_V^2, and σ_R^2. We can further replace σ_U^2 with $\sigma_U^2 = e^{\tau_U}$ and similarly for σ_V^2 and σ_R^2, to impose the constraint that variances are positive. See Supplement B for details on these moments.

Higher moments of U, V, and R can be estimated analogously. Alternatively, as discussed later, once we have identified and estimated β and γ, we can apply Kotlarski's Lemma to recover the entire distributions of U, V, and R.

We can also easily extend this identification and associated estimation to allow for covariates. Suppose we have the model

$$Y = b_1'X + U + V \tag{15}$$

$$W = \gamma Y + b_2' X + \beta U + R \tag{16}$$

where X is exogenous and is therefore uncorrelated with U, V, and R. The reduced form for W is now

$$W = (\gamma b_1 + b_2)' X + (\gamma + \beta) U + \gamma V + R$$

So we can estimate the coefficient vectors b_1 and b_2 along with γ and β by replacing Y and W in equations (9), (10), and (11) with $Y - b_1' X$ and $W - (\gamma b_1 + b_2)' X$, respectively and estimate those moments along with the moments

$$E\left(\left(W - (\gamma b_1 + b_2)' X\right) X\right) = 0, \qquad E\left((Y - b_1' X) X\right) = 0 \tag{17}$$

The complete set of moments for estimating this model via GMM, which we use in our empirical application, is provided in Supplement B.

Although we did not find this to be the case in our application, when GMM models are substantially overidentified (many more moments then parameters) it is sometimes preferable to only use a subset of available moments for estimation. Since our estimator takes the form of standard GMM, in these cases the existing literature on empirical choice of moments in standard GMM estimation might be applied. See, e.g., Andrews and Lu (2001), Caner (2009), and Liao (2013).

For simplicity, these estimators assumed the errors U, V, and R are homoskedastic, and similarly have higher moments that do not depend on X. This could be relaxed to allow higher moments of these errors to depend in unknown ways on X, by letting the assumptions of Lemma 1 hold conditional on X, thereby replacing the unconditional moments of equations (7) and (8) with conditional moments. Corresponding estimators would then, however, be much more complicated, and parameters like the error variances would need to be replaced by nonparametric functions of X.

3 General Point Identification

We now provide a more general and systematic analysis of the identification of our model, using more information than the low order moments of Lemma 1. We provide four main

results. First, we show that it is possible to construct infinitely many moments like those of Lemma 1, which can be used to construct simple GMM estimators, and we give the conditions under which these moments point identify the coefficients α and γ (equivalently, β and γ). Second, we apply Kotlarski's lemma to point identify the distributions of U, V, and R given point identification of α and γ. Third, we demonstrate that, using the entire joint distribution of Y and W (instead of just some moments) the only case where point identification is not possible is when U or V (or both) are normal. Finally, in the not point identified case, we fully characterize the sharp identified set.

We make extensive use of the characteristic function and its logarithm. Knowing the (log) characteristic function of a vector of random variables is equivalent to knowing the joint distribution of those variables (Theorem 3.1.1 in Lukacs (1970)).

Definition 1 *Given two random variables Y and W, let $\phi_{Y,W}(\zeta,\xi) \equiv E\left[e^{i\zeta Y+i\xi W}\right]$ denote their joint characteristic function. Similarly for a single random variable, let $\phi_Y(\zeta) \equiv E\left[e^{i\zeta Y}\right]$. Moreover, let $\Phi_{Y,W}(\zeta,\xi) \equiv \ln \phi_{Y,W}(\zeta,\xi)$ and $\Phi_Y(\zeta) \equiv \ln \phi_Y(\zeta)$ denote log characteristic functions (which are also called cumulant generating functions).*

Definition 2 *Given two random variables Y and W, define the cumulant of order k,ℓ (Lukacs (1970), p. 27) as*

$$\Phi_{Y,W}^{k,\ell} \equiv \left[\frac{\partial^{k+\ell}\Phi_{Y,W}(\zeta,\xi)}{i^{k+\ell}\partial\zeta^k\partial\xi^\ell}\right]_{\zeta=0,\xi=0}.$$

Similarly for a single random variable, define the cumulant of order k as

$$\Phi_Y^k \equiv \left[\frac{\partial^k\Phi_Y(\zeta)}{i^k\partial\zeta^k}\right]_{\zeta=0}.$$

All cumulants can be expressed in terms of standard moments, as obtained by an explicit differentiation of the log characteristic function and by exploiting the characteristic function moment theorem (e.g. $E\left[Y^k\right] = \left[\frac{\partial^k\phi(\xi)}{i^k\partial\xi^k}\right]_{\xi=0}$)[7]. Also note that the joint and marginal

[7]For high-order cumulants, these otherwise tedious algebraic manipulations could be handled with symbolic algebra packages.

characteristic functions as well as the corresponding cumulants are directly related, e.g., $\phi_Y(\zeta) = \phi_{Y,W}(\zeta, 0)$, $\Phi_Y(\zeta) = \Phi_{Y,W}(\zeta, 0)$ and $\Phi_Y^k = \Phi_{Y,W}^{k,0}$.

With these tools in hand, we are ready to state a general identification result based on moment constraints. As in Lemma 1, we start by rewriting the model of equations (4) and (5) in the reduced form of equations (4) and (6), and focus on the parameters α and γ.

Theorem 1 *Let Assumptions 1, 2, and Equations (4) and (6) hold. Assume* $-\infty < \gamma < \alpha < \infty$ *and let*

$$M_p(\alpha, \gamma) \equiv \Phi_{Y,W}^{1+p,2} - \alpha^2 \Phi_Y^{3+p} - (\gamma + \alpha)\left(\Phi_{Y,W}^{2+p,1} - \alpha \Phi_Y^{3+p}\right). \tag{18}$$

Let $q, \tilde{q} \in \mathbb{N} \equiv \{0, 1, \ldots\}$ *with* $q < \tilde{q}$. *If* $E\left[|U|^{\tilde{q}}\right]$, $E\left[|V|^{\tilde{q}}\right]$ *and* $E\left[|R|^{\tilde{q}}\right]$ *exist and* $\Phi_Y^{3+\tilde{q}} \Phi_{Y,W}^{2+q,1} \neq \Phi_Y^{3+q} \Phi_{Y,W}^{2+\tilde{q},1}$ *(or, equivalently, if* $\Phi_U^{3+\tilde{q}} \Phi_V^{3+q} \neq \Phi_V^{3+\tilde{q}} \Phi_U^{3+q}$*), then the moment constraints*

$$M_q(\alpha, \gamma) = 0 \tag{19}$$

$$M_{\tilde{q}}(\alpha, \gamma) = 0 \tag{20}$$

point identify the parameters of the model as $(\alpha, \gamma) = (\alpha_+, \alpha_-)$*, where*

$$\alpha_\pm = \frac{F^{3012}}{2F^{3021}} \pm \sqrt{\left(\frac{F^{3012}}{2F^{3021}}\right)^2 + \frac{F^{1221}}{F^{3021}}}$$

and where $F^{abcd} \equiv \Phi_{Y,W}^{a+\tilde{q},b} \Phi_{Y,W}^{c+q,d} - \Phi_{Y,W}^{a+q,b} \Phi_{Y,W}^{c+\tilde{q},d}$.

The proof, provided in Supplement A, proceeds by a judicious choice of cumulants of (Y, W) that do not depend on cumulants of R, and by exploiting the fact that cumulants of (Y, W) of order k, ℓ that share the same value of $k + \ell$ involve the same cumulants of U and V with prefactors that only differ in how they depend on α and γ. These observations then lead to specific functions of cumulants that can be analytically solved for α and γ.

Note that Theorem 1 also relies on Assumption 4, here rephrased as $-\infty < \gamma < \alpha < \infty$. Had we assumed $-\infty < \alpha < \gamma < \infty$ instead, then essentially the same Theorem would hold except that now α and γ would be point identified by $(\alpha, \gamma) = (\alpha_-, \alpha_+)$. We next formally show that Theorem 1 contains Lemma 1 as a special case.

Corollary 2 *The assumptions of Theorem 1 with* $q = 0$ *and* $\tilde{q} = 1$ *imply that the assumptions of Lemma 1 hold. Equations (19) and (20) in Theorem 1 with* $q = 0$ *and* $\tilde{q} = 1$ *are equivalent to equations (7) and (8) in Lemma 1.*

Equations (9), (10), and (11), used for GMM estimation of α and γ, were obtained by converting equations (7) and (8) into moments suitable for GMM. Equivalently, equations (9), (10), and (11) could have been directly derived from $M_0(\alpha, \gamma) = 0$ and $M_1(\alpha, \gamma) = 0$. This is done explicitly in the proof of Corollary 2.

As noted above, all cumulants can be expressed in terms of standard moments, specifically, cumulants equal sums of products of moments. To fit within a GMM framework, the cumulants in the expressions $M_p(\alpha, \gamma) = 0$, after being converted to functions of moments, must be linearized. This is done by introducing nuisance parameters. To illustrate, the cumulant Φ_Y^4 appears in the equation $M_1(\alpha, \gamma) = 0$. Now Φ_Y^4 equals $E[Y^4] - 3[E(Y^2)]^2$, so, e.g., to convert the expression $\Phi_Y^4 = c$ into a form suitable for GMM, we rewrite this expression as $E[Y^4 - 3Y^2\mu_{YY} - c] = 0$ and $E[Y^2 - \mu_{YY}] = 0$, using the nuisance parameter μ_{YY} that was introduced in the previous section.

Theorem 1 shows that one can obtain any number of additional, potentially overidentifying, moments to use for GMM estimation, based on the fact $M_p(\alpha, \gamma) = 0$ holds for any nonnegative integer p (as long as the associated moments of U, V, and R exist). We illustrate this in Supplement B, where, in addition to the moments based on Lemma 1, we provide the additional moments suitable for GMM estimation that are obtained from $p = 2$. In our later Monte Carlo simulations and empirical application, we provide results using the exactly identifying set of GMM moments based on $p = 0$ and 1, and also using the generally over identifying set of GMM moments based on $p = 0, 1$ and 2.

Theorem 1 provides explicit conditions under which any pair of cumulant functions $M_q(\alpha, \gamma) = 0$ and $M_{\tilde{q}}(\alpha, \gamma) = 0$ suffice to identify the parameters α and γ. In particular, point identification based on the moments in Lemma 1, corresponding to $M_0(\alpha, \gamma) = 0$

and $M_1(\alpha,\gamma)=0$, requires that $\Phi_U^4\Phi_V^3 \neq \Phi_V^4\Phi_U^3$, or equivalently

$$\left(E\left(U^4\right)-3\left[E\left(U^2\right)\right]^2\right)E\left(V^3\right)-\left(E\left(V^4\right)-3\left[E\left(V^2\right)\right]^2\right)E\left(U^3\right)\neq 0. \tag{21}$$

The left-hand side of (21) turns out to be proportional to the determinant of the Jacobian of the moment conditions (7) and (8) evaluated at the true value of the parameters:

$$\beta\begin{bmatrix} E[V^3] & -E[U^3] \\ E[V^4]-3\left(E[V^2]\right)^2 & -E[U^4]+3\left(E[U^2]\right)^2 \end{bmatrix}. \tag{22}$$

This connection is expected, since having a nonsingular Jacobian at the true parameter values is a necessary condition for point identification.

Condition (21) is violated, for instance, if either U or V is normal, or if both U and V are symmetric, or if both U and V have the exact same distribution. If we add the additional moments corresponding to $M_2(\alpha,\gamma)=0$, then point identification only requires that at least one of the inequalities $\Phi_U^4\Phi_V^3 \neq \Phi_V^4\Phi_U^3$, $\Phi_U^5\Phi_V^3 \neq \Phi_V^5\Phi_U^3$, or $\Phi_U^5\Phi_V^4 \neq \Phi_V^5\Phi_U^4$, hold. For example, if the second of these holds then Theorem 1 applies with $q=0$ and $\widetilde{q}=2$. If more than one of these inequalities holds, then we are generally overidentified.

Once the parameters α and γ have been identified, the full distribution of all unobservables can be determined under the following Assumption.[8]

Assumption 5 *The characteristic functions of U, V and R are nonvanishing on the real line.*

Corollary 3 *If Assumptions 1, 2, 5 and Equations (4) and (6) hold, $E[|Y|]<\infty$ and if α,γ are point identified, then the distributions of U, V and R are point identified from the joint distribution of Y and W through*

$$\begin{aligned} \Phi_V(\xi) &= \int_0^{\xi} \frac{E\left[iYe^{i\zeta\frac{W-\alpha Y}{\gamma-\alpha}}\right]}{E\left[e^{i\zeta\frac{W-\alpha Y}{\gamma-\alpha}}\right]}d\zeta \\ \Phi_U(\zeta) &= \Phi_Y(\zeta)-\Phi_V(\zeta) \\ \Phi_R(\xi) &= \Phi_W(\xi)-\Phi_U(\alpha\xi)-\Phi_V(\gamma\xi). \end{aligned} \tag{23}$$

[8]This can be relaxed to nonvanishing everywhere, except at isolated points, under slightly stronger moment existence conditions; see Schennach (2000) and Evdokimov, K. and H. White (2012).

A more explicit expression for the distributions of these unobserved variables can be obtained by an inverse Fourier transform. For instance, if V admits a density, it is given by

$$f_V(v) = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp(\Phi_V(\xi)) e^{-i\xi v} d\xi \tag{24}$$

and similarly for the other densities. More general distributions (e.g. discrete and/or singular) can be recovered as well, if equation (24) is interpreted in the appropriate measure theoretic sense.

Although Theorem 1 is quite general, it does require the condition $\Phi_U^{3+\tilde{q}}\Phi_V^{3+q} \neq \Phi_V^{3+\tilde{q}}\Phi_U^{3+q}$ to deliver identification, so it is natural to ask whether this is fundamentally necessary. It is in fact possible to formulate an estimation strategy that relaxes this condition. For instance, as discussed above, one could stack the moment conditions of the form (19) and (20) obtained with different values of $(q, \tilde{q})$. The resulting moment conditions would only fail to identify (α, γ) if the condition $\Phi_U^{3+\tilde{q}}\Phi_V^{3+q} \neq \Phi_V^{3+\tilde{q}}\Phi_U^{3+q}$ fails simultaneously for all the choices of q and $\tilde{q}$ considered.

An even more general strategy could be to start from the fundamental relationships between the log characteristic functions of the observables and unobservables ($\Phi_{Y,W}(\zeta, \xi) = \Phi_U(\zeta + \alpha\xi) + \Phi_V(\zeta + \gamma\xi) + \Phi_R(\xi)$) and cast identification as an optimization problem that minimizes deviations between the observed quantities (i.e. $\Phi_{Y,W}(\zeta, \xi)$) and predicted quantities:

$$\begin{aligned} & (\alpha, \gamma, \Phi_U, \Phi_V, \Phi_R) \\ = \ & \arg \min_{(\alpha,\gamma,\Phi_U,\Phi_V,\Phi_R)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\Phi_U(\zeta + \alpha\xi) + \Phi_V(\zeta + \gamma\xi) + \Phi_R(\xi) - \Phi_{Y,W}(\zeta, \xi)\right|^2 d\xi d\zeta, \end{aligned} \tag{25}$$

subject to $\alpha \geq \gamma$, zero mean constraints ($\Phi_U'(0) = 0, \Phi_V'(0) = 0, \Phi_R'(0) = 0$) and that (Φ_U, Φ_V, Φ_R) be valid log characteristic functions. This approach circumvents requiring existence of the moments $E\left[|U|^{\tilde{q}}\right]$, $E\left[|V|^{\tilde{q}}\right]$ and $E\left[|R|^{\tilde{q}}\right]$. However, the introduction of nuisance functions (Φ_U, Φ_V, Φ_R) would complicate estimation, as these would have to be parameterized by series or other expansions to construct a corresponding sieve estimator.

An estimator based on Equation (25) would be obtained replacing $\Phi_{Y,W}(\zeta,\xi)$ by its sample analogue and trimming or downweighting the high-frequency tails in the integral.

The question remains, do there exist situations where neither this nor any other estimator can consistently estimate the model, due to lack of point identification? The following theorem fully addresses this question, by showing that there exist cases that are not point identified. However, all such cases are when U or V (or both) are normal.

This differs from, and is simpler than, Reiersøl's (1950) well-known result in linear univariate errors-in-variables models, where the nonidentified cases arise when the model contains normal factors (see below). However, the required methods of proof differ significantly. For instance, the presence of two slope parameters α and γ (instead of one), and the presence of both latent variables U and V in both equations of the model, prevents us from using Reiersøl's proof method, which is based on the fact that two functions of different variables that are equal to each other must be constant. In our case, we have sums of many different functions of different variables on each side of an equality, and possible cancellation between terms that complicates the argument significantly.

Assumption 6 $E\left[|U|^3\right], E\left[|V|^3\right], E\left[|R|^3\right]$ *are finite.*

Theorem 4 *Let Assumptions 1, 2, 5, 6 and Equations (4) and (6) hold and assume that* $-\infty < \gamma < \alpha < \infty$. *If neither* U *nor* V *are normally distributed, then* α, γ *are uniquely determined by the joint distribution of* Y *and* W *by Equation (25).*

Note that U or V normal implies Y has full real line support, so having the support of Y be bounded is a simple sufficient condition for point identification. In the next section, we address what happens when either U or V (or both) are normally distributed.

4 Set Identification

In the case where Theorem 4 does not apply, so that the parameters are not point identified, the objective function of Equation (25) is maximized over a set rather than at a

single point. In order to precisely characterize this *identified set*, we first need to introduce the notion of *factor*, which is used by Reiersøl (1950) and by Schennach and Hu (2013).

Definition 3 *If a random variable Z can be decomposed as $Z = Z_1 + Z_2$ where Z_1 and Z_2 are independent, then Z_1 and Z_2 are called* factors *of Z. (The term* factor *can also be used to refer to the distributions of these variables.)*

While for given characteristic functions $\phi_{Z_1}(\xi)$ and $\phi_{Z_2}(\xi)$, we automatically have that $\phi_Z(\xi) = \phi_{Z_1}(\xi)\,\phi_{Z_2}(\xi)$ by the convolution theorem, the notion of factor embodies the fact that, if one is instead given the two characteristic functions $\phi_Z(\xi)$ and $\phi_{Z_1}(\xi)$, it is not automatic that there exists a random variable Z_2 with characteristic function $\phi_{Z_2}(\xi) = \phi_Z(\xi)/\phi_{Z_1}(\xi)$. The inverse Fourier transform of $\phi_{Z_2}(\xi)$, may not actually yield a proper probability measure (it could assign negative weights to some sets, for instance).

Next we consider what it means for a random variable to have a normal factor.

Lemma 2 *Let Z be an observed zero mean random vector. Then Z admits a unique decomposition into two unobserved zero mean independent factors*

$$Z = Z_g + Z_n, \tag{26}$$

where Z_g is Gaussian with variance $\bar{\Lambda}$ and Z_n has no Gaussian factors. Furthermore, the variance of Z_g is determined (from the observed distribution of Z) from the unique $\bar{\Lambda}$ such that

$$\bar{\Lambda} - \Lambda \text{ is positive semidefinite} \iff \phi_Z(\xi)\exp(\xi'\Lambda\xi/2) \text{ is a characteristic function.}$$

(Note that either Z_g or Z_n or both could be zero.)

Intuitively, Lemma 2 indicates that the decomposition into a Gaussian and a non-Gaussian factor can, in principle, be found by attempting to deconvolve Z by a Gaussian of variance Λ and seeking the "largest" (in a positive definite sense) possible Λ that will still yield a proper distribution. In Fourier representation, this amounts to dividing $\phi_Z(\xi)$ by

$\exp(-\xi'\Lambda\xi/2)$ and checking if the result is a valid characteristic function (e.g., by verifying if the inverse Fourier transform is a nonnegative measure). An alternative check for the validity of a given function $\phi(\xi)$ to be a valid characteristic function can be based on Bochner's Theorem (Theorem 4.2.2 in Lukacs (1970)): ϕ is a characteristic function iff

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j^* \phi(\xi_i - \xi_j) \geq 0 \text{ for all } c_1,\ldots,c_n \in \mathbb{C} \text{ for all } \xi_1,\ldots,\xi_n \in \mathbb{R} \text{ for all integer } n \geq 1$$

(Bochner's Theorem also includes the conditions that $\phi(\xi)$ be continuous and $\phi(0) = 1$ but these are automatically satisfied in our context.)

Using Lemma 2, we can decompose the observed $Z = (Y, W)$ into Gaussian (g) and non-Gaussian (n) factors

$$(Y, W) = (Y_g, W_g) + (Y_n, W_n) \tag{27}$$

This decomposition can be accomplished without the knowledge of α or γ. The non-Gaussian or Gaussian nature of the two factors is important in our context, because it is associated with the features that can or cannot be point-identified. This type of decomposition is not a purely theoretical construct; it can be empirically implemented. Independent Component Analysis techniques, which are widely used in signal processing, (see Hyvärinen and Oja (2000) for a review) specifically rely on such decompositions into Gaussian and non-Gaussian components.

Define

$$B_s = \frac{E[W_s Y_s]}{E[Y_s^2]} \tag{28}$$

$$D_s = \frac{E[W_s^2]\,E[Y_s^2] - (E[W_s Y_s])^2}{(E[Y_s^2])^2} \geq 0 \tag{29}$$

where the subscript s is either set to "g" , or to "n" , or is removed. We can now state our set-identification theorem:

Theorem 5 *Let Assumptions 1, 2 and Equations (4) and (6) hold and assume that* $E[Y^2]$, $E[W^2]$, $E[R^2] < \infty$ *and that* $-\infty < \gamma < \alpha < \infty$. *Then, the following bounds (illustrated in Figure 1) are sharp:*

1. If both U and V are Gaussian (and $E[Y^2] > 0$), then

$$\begin{aligned} \alpha &\geq B_g & (30) \\ B_g - \frac{D_g}{\alpha - B_g} &\leq \gamma \leq B_g. & (31) \end{aligned}$$

2. If V is Gaussian but U is not (and $E[Y_n^2]$, $E\left[Y_g^2\right] > 0$), then

$$\begin{aligned} \alpha &= B_n & (32) \\ B_g - \frac{D_g}{\alpha - B_g} &\leq \gamma \leq B_g. & (33) \end{aligned}$$

3. If U is Gaussian but V is not (and $E[Y_n^2]$, $E\left[Y_g^2\right] > 0$), then

$$\begin{aligned} \gamma &= B_n & (34) \\ B_g &\leq \alpha \leq B_g + \frac{D_g}{B_g - \gamma}. & (35) \end{aligned}$$

For each of the possible values of (α, γ) in the set given by Theorem 5, there corresponds a unique implied distribution for U, for V, and for R, given by Corollary 3. To distinguish between the three cases in Theorem 5, we have that case 1 holds only if Y is normal, in case 2 $B_n > B$, and in case 3 $B_n < B$.

Although the quantities B_n, B_g, D_n, D_g are, in principle, observable quantities, they may be difficult to estimate. For this reason, we also provide below a coarser bound that is only based on the covariances matrix of the observed Y and W:

Corollary 6 *The following bounds on α, γ always hold:*

$$\begin{aligned} \alpha &\geq B \\ B - \frac{D}{\alpha - B} &\leq \gamma \leq B. \end{aligned}$$

It is no accident that these bounds have the same form as Case 1 of Theorem 5: Both are solely based on covariance information, but in the Gaussian case, covariances exhaust all available information and yield sharp inequalities while, in general, that is not the case.

This looser bound is also related to the measurement error bounds in Frisch (1934). If one is willing to rely on this relaxed bound, then a simple GMM estimator for the resulting identified set could be obtained based on the moment conditions

$$E\left[\alpha^2\sigma_U^2 + \gamma^2\left(Y^2 - \sigma_U^2\right) + \sigma_R^2 - W^2\right] = 0 \tag{36}$$

$$E\left[\alpha\sigma_U^2 + \gamma\left(Y^2 - \sigma_U^2\right) - YW\right] = 0 \tag{37}$$

while optimizing over $\alpha, \gamma, \sigma_U^2, \sigma_R^2$, subject to the constraints $\gamma < \alpha$ (equivalent to $\beta > 0$), $\sigma_U^2 \geq 0$ and $\sigma_R^2 \geq 0$. These moment conditions are obtained from Equations (66) and (67) in the proof of Theorem 5, without extracting the Gaussian parts. The bounds of Corollary 6 are also obeyed in the case of point identified models, since they are obtained solely from positive variance considerations that must always be satisfied. This implies that, if one is unsure whether Y is normal or not, the moment conditions (36) and (37) could be stacked with the ones of Theorem 1 to yield an estimator that is robust to loss of point identification.[9]

5 Ordinary Least Squares

It is instructive to analyze in more detail how the parameters of our model relate to the slope coefficient of a naive OLS regression (in the population limit). The coefficient B given by Equation (28) is the slope coefficient of the least-square regression of W on Y (in the population limit). Regardless of whether the model is point identified or not, an implication of the model (i.e., of equations (4) and (5)) is that B always lies between γ and α. This can be immediately verified by observing that

$$B = \frac{E[YW]}{E[Y^2]} = \frac{E[(U+V)(\alpha U + \gamma V)]}{E\left[(U+V)^2\right]} = \frac{\alpha E[U^2] + \gamma E[V^2]}{E[U^2] + E[V^2]} = \alpha\lambda + \gamma(1-\lambda) \tag{38}$$

where $\lambda = E[U^2] / (E[U^2] + E[V^2])$ and so lies between zero and one. So in particular, if $\beta > 0$ we get $\gamma \leq B \leq \alpha$.

This type of inequality has been noted before in the context of estimating returns to education (e.g. by Card (2001), in a more detailed model that allows for some individual

[9] In this case the maximizing estimands could be sets rather than points, requiring nonstandard inference.

heterogeneity). In particular, in the returns to schooling context, we would expect both β and γ to be positive (because unobserved ability U should affect schooling Y and wages W in the same direction, and increased schooling should increase wages). By the above analysis, this in turn means that we would expect $0 < \gamma \leq B$.

However, as noted by Card (2001), most returns to schooling empirical applications yield estimates of γ, using instrumental variables methods, that are greater than B, which contradicts this inequality and hence also contradicts the model. One possible explanation for this contradiction is that, in the returns to schooling context, Y may also contain significant measurement error. Standard attenuation bias under classical measurement error implies that the ordinary least squares coefficient B is biased towards zero relative to γ, which if $0 < \gamma$ would imply $B < \gamma$. If the model is correct for returns to education, but in addition Y is mismeasured, then B could be either larger or smaller than γ, depending on the relative magnitude of the measurement error.

6 Monte Carlo

To assess the finite sample performance of our simple GMM estimators, we generate data from the model of equations (4) and (5) without covariates. All of our designs are chosen to satisfy equation (21), so the model is point identified just from the moments in Lemma 1.[10] The true values of the coefficients are $\gamma = \beta = 1$. It is widely recognized that estimators based on higher moments can behave poorly with small sample sizes, so to see if our estimators suffer from these issues, we work with relatively small sample sizes of $n = 100$ and $n = 400$.

We generate $5,000$ replications of four different designs. In design 1, U is log normal while V and R are each standard Gumbel. In design 2, U is log normal while V and R are uniform. We then reverse these, making U Gumbel and V and R log normal in design 3, and making U uniform with V and R log normal in design 4. For each design, we report results using two different estimators. The exactly identified estimator is GMM using moments

[10] In particular in all of our designs, U and V have different, non-normal distributions, and at least one is asymmetically distributed. U, V, and R are also mutually independent and centered at mean zero.

corresponding to Lemma 1, given by equations (77), (78), and (79) (without covariates, so $\widetilde{Y} = Y$ and $\widetilde{W} = W$), as given in Supplement B. The over-identified estimator is GMM using these same equations, plus equations (81) and (82) of Supplement B.

Tables C1 to C4 of the Supplement report results from designs 1 to 4, respectively. Each Table has four panels, corresponding to the two different GMM estimators, each with the two different sample sizes. We report estimates of γ, β, the error component variances σ_U^2, σ_V^2, and σ_R^2, and, when over-identified, μ_{WW}. Reported summary statistics of each parameter estimate across the simulations are the mean (MEAN), the standard deviation (SD), the 25% quantile (LQ), the median (MED), the 75% quantile (UQ), the root mean squared error (RMSE), the mean absolute error (MAE), and the median absolute error (MDAE).

Some general tendencies stand out in these simulations. First, consider the trade off between the exactly identified vs over identified estimators. The latter uses more information, but that information takes the form of up to fifth order moments, which can be noisy and more sensitive to outliers. In general we find that the overidentified estimator performs better than the exactly identified estimator, particularly at the larger sample size.

The primary parameter of interest, γ, tends to be estimated reasonably precisely in all of the designs, with most RMSEs in the range of .3 to .7. In contrast, β is generally much less precisely estimated, often having much larger RMSEs (except in design 2). Estimates of the variances σ_U^2, σ_V^2, and σ_R^2, are mostly similar to each other, usually being less precise than γ but more than β. The estimate of μ_{WW} is noisier, since it only appears in the highest order moment equations of the over identified model. The designs where U was log normal (designs 1 and 2) generally had more accurate estimates than the other designs. We conclude that our estimator performs reasonably well even with rather small sample sizes.

7 GDP and Life Expectancy

There is a long literature studying the causal effect of health on economic growth. Examples include Acemoglu and Johnson (2007) (which we will hereafter refer to as AJ),

Well (2007), Lorentzen, McMillan, and Wacziarg (2008), Aghion, Howitt, and Murtin (2010), Cervellati and Sunde (2011), Ecevit (2013), Bloom, Canning, and Fink (2014), and Bloom, Canning, Kotschy, Prettner, and Schünemann (2019).

Based on a neo-classical growth model, AJ estimate a model in the form of equations (1) and (2), where Y is the change in the log of a country's life expectancy at birth between 1940 and 1980, W is the change in that country's log GDP in the same time span, and X is either just a constant, or a constant and a measure of the country's quality of institutions, or a constant and GDP per capita in 1930. The main goal is estimation of γ, the coefficient of Y in the W equation.

AJ observe that ordinary least squares estimation of the W equation is inconsistent, because the health measure Y is endogenous, with improvements and investments in a country's productive technology over time positively impacting both health outcomes and GDP. This technology change corresponds to our unobserved factor U (with $\beta > 0$) in equations (15) and (16), while V and R are the idiosyncratic shocks to health and economic outcomes, respectively.

To deal with the endogeneity caused by U, AJ construct an instrument, called predicted mortality, that combines each country's 1940 mortality rates from specific diseases with a set of global interventions that addressed those diseases. As noted in the introduction, one may question the validity of such constructed instruments.

In Table 1, columns labeled 2SLS1, 2SLS2, and 2SLS3 in Panel A are replications of selected results appearing in Table 9 of AJ.[11] These are AJ's estimates using two stage least squares (2SLS) with the above listed combinations of covariates X, and using their predicted mortality instrument. AJ's ordinary least squares (OLS) estimate of γ (corresponding to B in the previous section) is -0.81, while their 2SLS estimates of γ are considerably larger in magnitude, ranging from -1.316 to -1.643. As we noted earlier, having $\gamma < B$, as AJ find, is an implication of our model when $\beta > 0$. Note that the sample size is quite small in

[11] Our data are provided by AJ. Life expectancy is from UN data sources and the League of Nation reports. Pre-war GDP data are from Maddison (2003), and post-war data are from the UN. See AJ for details.

this application, with only 47 countries. Nevertheless, AJ's estimates of γ are statistically significant.[12]

Now suppose we had not observed predicted mortality, or we are uncertain of its validity as an instrument. We can instead consider applying our GMM estimators. First, consider the distribution of Y. Assuming (measured) life expectancy is bounded away from zero, log life expectancy is bounded, which suffices for point identification since it rules out U or V being normal.[13] We therefore attempt to apply our GMM estimators.

In Table 1, we report two sets of GMM estimates along with AJ's 2SLS results. Columns labeled GMM1, GMM2, and GMM3 are GMM estimates of equations (15) and (16), which do not make use of the predicted mortality instrument in any way. Specifically, these are estimates based on the over-identifying set of moments given by equations (77) to (82) in Supplement B. The last three columns of Table 1 then give GMM estimates that use both our over-identifying set of moments and the additional moment given by AJ's instrument (as discussed at the end of Supplement B).[14]

Panel A in Table 1 reports the main parameter of interest γ, and also reports b_2, the other covariate coefficients in equation (16). The variables in columns (4) and (7) have been demeaned so there is no constant.[15] Our main takeaway from Panel A of Table 1 is that our estimates of γ are quite comparable to AJ's. In GMM1 and GMM2, the estimates of γ are -1.984 and -1.241, virtually the same range as AJ's 2SLS estimates, and are

[12]Our standard errors in columns (1)-(3) of Table 1 differ from those reported by AJ. AJ's estimates are from *ivreg* in Stata 9. We use *ivregress 2sls*, which replaced *ivreg* as of Stata 10. *ivreg* and *ivregress* can give different robust standard error estimates, because *ivreg* uses HC1 (MacKinnon and White 1985) robust standard errors while *ivregress 2sls* uses HC0 (Huber-White). Also, to reduce the number of coefficients in GMM estimation, we differenced the data while AJ used level data with fixed effects. Since T=2, these are asymptotically equivalent estimators.

[13]More heuristically, if Y is close to normal, then it may be that U or V is close to normal. Y has a skewness of 0.170 and a kurtosis of 1.791, which is reasonably far from normal in terms of the low order moments our GMM estimator is based on. The p-value of a Shapiro-Wilk test of normality of Y is .02, rejecting normality, and even lower if one tests the residuals after regressing Y on either of the covariates in X.

[14]These GMM models are estimated in Stata, using the vce(robust) option to compute standard errors.

[15]In Supplement B: Moments for GMM Estimation, it is noted that "For the model without covariates, one can replace b_1 and b_2 with zero in the above expressions, and drop equation (80). Note that in this case Y and W should be demeaned." In columns (4) and (7), we demeaned Y and W so b_1 and b_2 are zeros.

statisically significant. GMM3 gives an estimate of a lower magnitude -0.383, but this estimate is statistically insignificant with a very large standard error, suggesting that our higher moment based estimator is imprecise for this particular combination of covariates and small sample size. The last three columns of Table 1, which combine both our moments and the AJ instrument, give estimates very close to those of AJ, with somewhat smaller standard errors, which is exactly what one would expect to see if both sets of moments are valid and if AJ's instrument is strong. In the bottom row of Table 1 we report Hansen's J-test; we do not reject validity of the joint set of overidentifying restrictions in any of the GMM estimates.

Panels B and C of Table 1 provide the other estimated parameters of the model. Panel C gives the estimated b_1 coefficients from equation (15), while Panel B gives the estimates of β and the estimated variances of our error components. β appears to be difficult to precisely estimate, with large standard errors.[16] In the specifications where γ is statistically significant, the variance of U (the source of endogeneity in the model) is much smaller than the variances of the idiosyncratic components V and R, but very precisely estimated with small standard errors.

Later tables have the same format as Table 1, providing additional results. In Table 2, we re-estimate the model using the exactly identified set of moments from Lemma 1. As expected with fewer moments, these estimates are less efficient, and turn out to be quite a bit noisier than those of Table 1. GMM5, with the quality of institutions as the covariate, is still reasonably comparable to AJ with γ of -1.401, while now both GMM4 and GMM6 are insignificant and more variable. The estimates combining these moments with AJ's instrument behave as before.

We also perform a number of robustness checks in Supplement D, using alternative outcome variables that AJ considered in their Tables 8-9. These additional outcomes are log population, log births, percentage of population under age 20, log GDP, and log GDP per working age population. Some of the alternative outcomes suffer from the issue that U might

[16] In contrast α is, like γ, much more precisely estimated, but apparently the difference $\beta \equiv \alpha - \gamma$ is harder to pin down.

also contain measurement error, and in those cases, our identification results would not apply. The results of our GMM estimators with other outcomes are generally more erratic than with log per capita GDP. The estimates that combine our moments and the AJ instrument remain comparable to AJ's 2SLS estimates.

We conclude that, in all specifications where the standard errors were small enough to yield statistically significant results, our estimates based on higher moments, without side information, are very close to those obtained by AJ that required an instrument.

8 Conclusions

We have shown that a standard linear triangular structural model is generally point identified, without an instrument or other side information that is generally used to identify such models. We illustrate the result with Monte Carlo simulations and in an empirical application. Our application shows that, without using an instrument, GMM estimation of moments based on the model yields estimates close to those that were obtained by previous authors using an instrument. Even when instruments are available, our estimator could be usefully combined with instrument based moments to either increase estimation precision by adding more moments to the model, or to provide overidentifying moments that might be used for specification testing.

What makes point identification possible is the assumed error structure, which takes the standard form of a scalar common component U in each equation, plus additional scalar idiosyncratic components V and R. One goal for future work could include deriving alternative estimators for the model. These could include estimators that allow U, V, and R to depend nonparametrically on covariates X (e.g., allowing heteroskedasticity of unknown form), and estimators that make direct use of all the information in Theorem 4, perhaps based directly on characteristic functions rather than moments. Other possibilities for further work include extending the model to more equations, allowing the common component U to affect outcomes nonlinearly, and extending the model to also allow for measurement

error in Y. Based on Card (2001), this last extension would likely be needed for returns to education applications.

Acknowledgments

Susanne Schennach acknowledges support from NSF grant SES-1950969. Vincent Starck is gratefully acknowledged for valuable comments.

Disclosures

The authors report there are no competing interests to declare.

References

Acemoglu, D., and S. Johnson, (2007), "Disease and development: The effect of life expectancy on economic growth," Journal of Political Economy, 115(6), 925-985.

Aghion, P., P. Howitt, P., and F. Murtin, (2010), "The relationship between health and growth: When Lucas meets Nelson-Phelps, " National Bureau of Economic Research.

Andrews, D.W.K. and Lu, B. (2001): "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models", Journal of Econometrics, 101(1), 123-164.

Angrist, J. and A. Krueger, (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" Quarterly Journal of Economics, 106, 979-1014.

Angrist, J. and A. Krueger, (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," Journal of Economic Perspectives, 15, 69-85.

Bierens, H. J. (1981) Robust Methods and Asymptotic Theory in Nonlinear Econometrics, Springer Verlag

Bloom, D. E., D. Canning, and G. Fink (2014), "Disease and development revisited ," Journal of Political Economy, 122(6), 1355-1366.

Bloom, D. E., D. Canning, R. Kotschy, K. Prettner, and J. J. Schünemann (2019), "Health and Economic Growth: Reconciling the Micro and Macro Evidence," NBER Working Paper No. 26003.

Bonhomme, S. and J. - M. Robin (2010), "Generalized Non-Parametric Deconvolution with an Application to Earnings Dynamics," The Review of Economic Studies, 77, 491–533.

Caner, M. (2009): "Lasso-type GMM Estimator", Econometric Theory, 25(1), 270-290.

Card, D (1995) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," In L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp , University of Toronto Press, Toronto.

Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," Econometrica, 69, 1127–1160.

Carneiro, P. and J. J. Heckman, (2002), "The Evidence on Credit Constraints in Post-Secondary Schooling," The Economic Journal, 112(482), 705–734.

Cervellati, M., and U. Sunde, (2011), "Life expectancy and economic growth: The role of the demographic transition,"Journal of economic growth, 16(2), 99-133.

Comon, P. (1994), "Independent component analysis, a new concept?," Signal processing, 36(3), 287-314.

Darmois, G. (1953), "Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire," Revue de l'Institut international de statistique, 2-8.

Fruehwirth, J. C., S. Navarro, and Y. Takahashi (2016), "How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects," Journal of Labor Economics 34:4, 979-1021

Erickson, T. and T. M. Whited, (2002), "Two-step GMM estimation of the errors-in-

variables model using high-order moments," Econometric Theory, 18(3), 776-799.

Evdokimov, K. and H. White (2012), "Some Extensions of a Lemma of Kotlarski," Econometric Theory, 28(4), 925–932.

Frisch, R. (1934), "Statistical confluence analysis by means of complete regression systems," Vol. 5, Universitetets Økonomiske Instituut.

Hyvärinen, A. and E. Oja (2000) "Independent component analysis: algorithms and applications," Neural Networks 13, 411–430.

Khatri, C. and Rao, C. R. (1972), "Functional equations and characterization of probability laws through linear functions of random variables," Journal of Multivariate Analysis 2, 162–173.

Klein, R., and F. Vella, (2010), "Estimating a class of triangular simultaneous equations models without exclusion restrictions," Journal of Econometrics 154(42), 154–164.

Kotlarski, I. I. (1967), "On characterizing the gamma and normal distribution," Pacific Journal of Mathematics, 20, 69–76.

Lewbel, A. (1997), "Constructing Instruments for Regressions With Measurement Error When No Additional Data are Available, With an Application to Patents and R&D," Econometrica, 65(5), 1201-1213.

Lewbel, A. (2012), "Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models," Journal of Business and Economic Statistics, 30, 67–80.

Lewbel, A. (2020), "Kotlarski With a Factor Loading," Unpublished Manuscript, Boston College.

Li, S. and X. Zheng, (2020), "A Generalization of Lemma 1 in Kotlarski (1967)," Statistics and Probability Letters,165, article 108814.

Li, T. and Q. Vuong (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," Journal of Multivariate Analysis, 65, 139–165.

Liao, Z. (2013): "Adaptive GMM Shrinkage Estimation With Consistent Moment Selection", Econometric Theory, 29(5), 857-904.

Lorentzen, P., J. McMillan, and R. Wacziarg, (2008), "Death and development," Journal of Economic Growth, 13(2), 81–124.

Lukacs, E. (1970), "Characteristic Functions," Second edition, Griffin, London.

Maddison, A. (2003), "Development centre studies the world economy historical statistics: Historical statistics," OECD Publishing.

Navarro, S. and J. Zhou, (2017), "Identifying agent's information sets: An application to a lifecycle model of schooling, consumption and labor supply," Review of Economic Dynamics, 25, 58-92.

Peters, J., D. Janzing, and B. Scholkopf, (2017), "Elements of causal inference: foundations and learning algorithms," MIT press.

Rao, C. R. (1966), "Characterisation of the distribution of random variables in linear structural relations," Sankhyā: The Indian Journal of Statistics, Series A, 251-260.

Rao, C. R. (1971), "Characterization of probability laws by linear functions," Sankhyā: The Indian Journal of Statistics, Series A, 265-270.

Reiersøl, O. (1950), "Identifiability of a linear relation between variables which are subject to error," Econometrica, 18, 375-389.

Rigobon, R. (2003), "Identification Through Heteroskedasticity," Review of Economics and Statistics 85(4), 777–792.

Schennach, S. M. (2000), "Estimation of nonlinear models with measurement error," Working Paper, University of Chicago.

Schennach, S. M. (2019), "Convolution without independence," Journal of Econometrics, 211(1), 308-318.

Schennach, S. M. and Y. Hu (2013), "Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models Without Side Information," Journal of

the American Statistical Association, 108, 177-186.

GJ Székely, G. J. and C. R. Rao (2000), "Identifiability of distributions of independent random variables by linear combinations and moments," Sankhyā: The Indian Journal of Statistics, Series A, 62(2), 193-202.

Well, D. N. (2007), "Accounting for the Effect Of Health on Economic Growth," The Quarterly Journal of Economics, 122(3), 1265–1306.

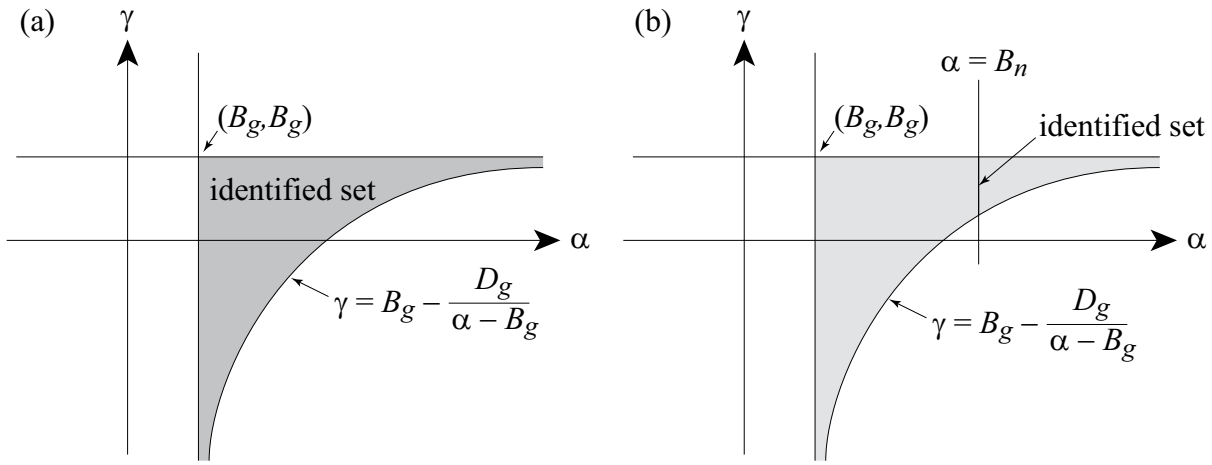

Figure 1: Identified set of Theorem 5 for (a) Case 1 and (b) Case 2 (Case 3, analogous to Case 2, is not shown).

Table 1: Over identified moments: Base sample 1940 and 1980

	(1) 2SLS1	(2) 2SLS2	(3) 2SLS3	(4) GMM1	(5) GMM2	(6) GMM3	(7) GMM1+AJ	(8) GMM2+AJ	(9) GMM3+AJ
Panel A. Dependent Variable: Growth in GDP per Capita 1940-1980									
Life expectancy	-1.316***	-1.643***	-1.589*	-1.984**	-1.241**	-0.383	-1.341***	-1.642***	-1.573*
	(0.382)	(0.521)	(0.876)	(0.911)	(0.626)	(0.383)	(0.334)	(0.520)	(0.866)
Institutions		-0.0490			-0.0291			-0.0489	
		(0.0418)			(0.0472)			(0.0417)	
Initial (1930) value of			-0.0730			0.149			-0.0638
dependent variable			(0.198)			(0.112)			(0.193)
Constant	1.336***	1.681***	1.990		1.448***	-0.127		1.680***	1.910
	(0.124)	(0.367)	(1.807)		(0.445)	(0.983)		(0.366)	(1.760)
Panel B. β and variances									
β				2.319***	4.527	54.28*	3.966**	9.523***	1.215
				(0.743)	(11.71)	(115.1)	(2.277)	(5.420)	(1.601)
σ_U^2				0.0147***	0.00171**	6.68e-07	0.00467***	0.00152***	0.0136***
				(0.0136)	(0.00486)	(0)	(0.00436)	(0.00192)	(0.0155)
σ_V^2				0.0150***	0.0177***	0.0139***	0.0260***	0.0179***	4.72e-05
				(0.0143)	(0.00552)	(0.00444)	(0.00516)	(0.00412)	(0.0155)
σ_R^2				0.0547***	0.0943**	0.120***	0.0586***	1.75e-09	0.123***
				(0.0383)	(0.0973)	(0.0263)	(0.0250)	(0)	(0.0319)
μ_{ww}				0.147***	0.143***	0.124***	0.137***	0.143***	0.125***
				(0.0261)	(0.0271)	(0.0206)	(0.0235)	(0.0270)	(0.0207)
Panel C. Dependent Variable: Growth in Life Expectancy 1940-1980									
Institutions		-0.0310***			-0.0496***			-0.0496***	
		(0.00755)			(0.00997)			(0.00996)	
Initial (1930) value of			-0.117***			-0.184***			-0.185***
dependent variable			(0.0310)			(0.0223)			(0.0212)
Constant		0.324***	1.122***		0.579***	1.757***		0.580***	1.760***
		(0.0595)	(0.267)		(0.0523)	(0.173)		(0.0522)	(0.165)
Observations		47	47	47	47	47	47	47	47
Hansen J				0.122	0	0.000493	0.850	0.000109	0.0598
p-val				0.727	1	0.982			

Notes: In all models, the endogenous regressor is the changes in log life expectancy between 1940 and 1980. 2SLS1 is the two-stage least squares regression of growth in GDP per capita on growth in life expectency, using predicted mortality as the instrument. 2SLS2 includes a measure of quality of institutions as exogenous covariate. 2SLS3 adds the initial (1930) value of log GDP per capita. GMM1-GMM3 are the same models as 2SLS1-2SLS3 estimated by GMM estimators based on over identified moments. GMM1-GMM3+AJ combine our over identified moments and the AJ moment, i.e. $E(IV\varepsilon) = 0$. The last row reports the p value of the J statistics under the null hypothesis that the overidentifying restrictions are valid.

Table 2: Exactly identified moments: Base sample 1940 and 1980

	(1) 2SLS1	(2) 2SLS2	(3) 2SLS3	(4) GMM4	(5) GMM5	(6) GMM6	(7) GMM4+AJ	(8) GMM5+AJ	(9) GMM6+AJ
Panel A. Dependent Variable: Growth in GDP per Capita 1940-1980									
Life expectancy	-1.316***	-1.643***	-1.589*	-3.636	-1.401	-0.340	-1.344***	-1.634***	-1.599*
	(0.382)	(0.521)	(0.876)	(4.184)	(9.818)	(1.022)	(0.348)	(0.520)	(0.871)
Institutions		-0.0490			-0.0370			-0.0478	
		(0.0418)			(0.483)			(0.0416)	
Initial (1930) value of			-0.0730			0.156			-0.0706
dependent variable			(0.198)			(0.195)			(0.194)
Constant	1.336***	1.681***	1.990		1.541	-0.195		1.670***	1.969
	(0.124)	(0.367)	(1.807)		(5.663)	(1.836)		(0.365)	(1.775)
Panel B. β and variances									
β				3.091	13.61	1.589	2.228	0.805	1.235
				(3.588)	(2,713)	(5.015)	(1.235)	(6.184)	(1.190)
σ_U^2				0.0278***	0.000712	8.72e-06	0.00804***	0.0177	0.0138***
				(0.0107)	(0.128)	(0.00959)	(0.00714)	(0.136)	(0.0123)
σ_V^2				0.00253	0.0187	0.0139***	0.0220***	0.00170	0.000140
				(0.00960)	(0.128)	(0.0133)	(0.00718)	(0.136)	(0.0100)
σ_R^2				0.101***	9.05e-07	0.123***	0.0863***	0.126***	0.123***
				(0.0307)	(28.70)	(0.0306)	(0.0198)	(0.0843)	(0.0317)
Panel C. Dependent Variable: Growth in Life Expectancy 1940-1980									
Institutions		-0.0310***			-0.0496***			-0.0494***	
		(0.00755)			(0.00997)			(0.00996)	
Initial (1930) value of			-0.117***			-0.184***			-0.184***
dependent variable			(0.0310)			(0.0223)			(0.0219)
Constant		0.324***	1.122***		0.579***	1.761***		0.578***	1.758***
		(0.0595)	(0.267)		(0.0522)	(0.173)		(0.0522)	(0.170)
Observations		47	47	47	47	47	47	47	47
Hansen J							2.039	0.00564	0.0289
p-val							0.153	0.940	0.865

Notes: In all models, the endogenous regressor is the changes in log life expectancy between 1940 and 1980. 2SLS1 is the two-stage least squares regression of growth in GDP per capita on growth in life expectency, using predicted mortality as the instrument. 2SLS2 includes a measure of quality of institutions as exogenous covariate. 2SLS3 adds the initial (1930) value of log GDP per capita. GMM4-GMM6 are the same models as 2SLS1-2SLS3 estimated by GMM estimators based on exactly identified moments. GMM4-GMM6+AJ combine our exactly identified moments and the AJ moment, i.e. $E(IV\varepsilon) = 0$. The last row reports the p value of the J statistics under the null hypothesis that the overidentifying restrictions are valid.