Title: Data Combination and Instrumental Variables in Linear Models

Authors and Affiliations:
Christopher Khawand
Department of Economics
Michigan State University
khawandc@msu.edu

Background / Context:

Instrumental variables (IV) methods allow for consistent estimation of causal effects, but suffer from poor finite-sample properties and data availability constraints. Bound, Jaeger, and Baker (1995) establish that estimation with weak instruments, even under a weak relationship between the instrument and the error term, can lead to large inconsistencies and finite sample bias. IV estimates also tend to have relatively large standard errors, often inhibiting the interpretability of differences between IV and non-IV point estimates. Lastly, instrumental variables' idiosyncratic nature reduces their availability in data sets alongside outcome and other variables of interest.

Most prior work on two-sample IV has exploited multiple data sets with the goal of attaining identification. Following the independent discovery of a similar method by Klevmarken (1982), Angrist and Krueger (1992) propose a two-sample instrumental variables estimator that allows for estimation of each covariance matrix composing the IV estimator with separate data sets from the same population, where both data sets have an observed instrument in common, but not the dependent and endogenous variables. Arellano and Meghir (1992) correspondingly propose a method equivalent to two-sample two-stage least squares (TS2SLS) to identify a model of labor supply. Under the assumption that the different samples utilized are drawn from the same population, these estimators identify parameters of interest consistently. Given its computational convenience and favorable asymptotic properties (Inoue and Solon, 2005), the TS2SLS estimator is a natural choice for instrumental variables estimation under data combination.

While data combination in this context can be seen as a second-best solution—reserved for when identification cannot be secured through a single sample—it has the potential to provide additional useful information to applied researchers in any scenario. Even when a parameter of interest is identified and consistently estimated with a single sample, data combination can be preferable for the mean squared error (MSE) of IV estimates of that parameter. For example, estimating the first-stage relationship between the endogenous regressor and the instrument(s) with a larger sample than would be possible with the primary data set alone can improve MSE. Because finite sample bias of the TS2SLS estimator depends on sampling error in first stage estimation, this leads to a reduction in bias and potentially an increase in efficiency of the IV estimate. Moreover, the use of an auxiliary data set can provide additional covariates affecting the outcome of interest, which can be used to increase precision. Incidentally, these additional covariates can also be used in evaluating the exogeneity of an instrument. Lastly, quality of measurement may differ between primary and auxiliary data sets, and a data set with better measures of the instrument may also be preferable to use to estimate the first stage.

Purpose / Objective / Research Question / Focus of Study:

This paper aims to explore the properties and potential applications of data combination, specifically through the lens of the TS2SLS estimator. The paper, in its final form, will demonstrate the finite sample properties of the TS2SLS estimator and provide guidelines to empirical researchers to identify when using auxiliary data through the TS2SLS estimator results in preferable estimates. This will be done analytically in a basic framework where feasible, but more general propositions will be argued through simulation evidence.

Significance / Novelty of study:

While econometric literature has addressed data combination problems thoroughly through both Generalized Method of Moments (GMM) and non-parametric frameworks, little has been done to outline the practical scenarios under which data combination results in preferable estimates in the context of instrumental variables. This paper aims to outline the potential gains of and provide guidelines for successful implementation of the TS2SLS estimator. Specifically, the potential finite sample bias reduction and efficiency gains in a variety of situations can help researchers better estimate causal effects and test the robustness of their estimates. The data combination cases outlined in the next section show some examples of how data combination methodology can improve valid causal inference.

Statistical, Measurement, or Econometric Model:

Consider a simple linear system with one endogenous variable and a well-behaved error term:

$$y_1 = y_2\beta_1 + x\beta_2 + \varepsilon_1$$
$$y_2 = z\gamma + \varepsilon_2$$
$$Cov(\varepsilon_1, \varepsilon_2) \neq 0$$
$$\mathrm{E}(\varepsilon_1|z) = 0$$
$$\mathrm{E}(\varepsilon_1^2|z) = \sigma^2$$

x and z are vectors of exogenous variables. A linear regression of y_2 on y_1 and x produces a biased and inconsistent estimate of β_1. A two-stage least squares procedure, in which fitted values are generated from a first stage regression of y_2 on z and x and then used as the instrument for y_2 in an IV regression of y_2 on y_1 and x, produces a consistent but biased estimate of β_1.

Assume there are two data sets containing relevant covariates for this model, with sample sizes N_1 and N_2, respectively. The data sets are random samples from the same population. Table 1 outlines some different stylized cases of data availability. For example, in case 1, there are N_1 observations containing covariate values $\{y_1, y_2, z\}$ in data set 1, and N_2 observations containing covariate values $\{y_2, z\}$in data set 2. In every case (with the exception of case 0), $\beta_1$is identified and can be consistently estimated with data set 1 alone. The next section will provide some hypotheses regarding the econometric properties of combined estimation with both data sets as compared to using data set 1 alone, and then provide some preliminary simulation evidence for those claims.

Usefulness / Applicability of Method:

Define $\hat{\beta}_{TS2SLS}$ as the estimator for $\beta_1$corresponding to the data combination procedure that will be suggested for each case. $\hat{\beta}_{SS,2SLS}$ is the 2SLS estimator using data set 1 alone. $\hat{\beta}_{FS,2SLS}$ is the estimator corresponding to using 2SLS in the "full-sample" hypothetical case in which all covariates contained in data sets 1 and 2 are observed in a single data set, with a sample size equal to the greater of N_1 and N_2. Each case has an individualized procedure in order to efficiently use information from both data sets that is outlined in the appendix. Case 0 pertains to the classic motivation for data combination (e.g., Angrist & Krueger, 1992), where additional data sources are necessary to estimate the relationship between the endogenous regressor and the instrument, and is only presented for contrast.

Case 1 identifies a scenario in which consistent estimates can be calculated with data set 1 alone, but combined estimation with data set 2 would result in smaller bias and more efficient estimates due to data set 2's larger sample size. Simulation results find a significant difference (t = 4.7) in bias, and the standard errors of $\hat{\beta}_{TS2SLS}$ were 1/10th the size of the standard errors of $\hat{\beta}_{SS,2SLS}$ on average.

Case 2 is similar to Case 1, except that data set 1 also has exogenous regressors x_1 affecting y_1(that may in general be correlated with z) which must be appropriately "partialled out" of the instrument via application of the Frisch-Waugh-Lovell theorem. Similar methods are followed for the remaining cases. The simulation results provide evidence for the prediction about the bias made earlier: the bias of $\hat{\beta}_{TS2SLS}$ is bounded above by $Bias[\hat{\beta}_{SS,2SLS}]$ and below by $Bias[\hat{\beta}_{FS,2SLS}]$, with significant differences from those bounds (t = 33.4 and t = 36.2, respectively). The simulation is inconclusive for efficiency predictions, either because of lack of power in the simulation, or because of the specific calibration.

In Case 3, if variation in z is random and in turn $Cov(z, x_2) = 0$, then using additional covariates in data set 2 in estimation of the first stage should result in more precise estimates of γ if $Cov(y_2, x_2) \neq 0$. Fitted values for data set 1 can then be generated as $\hat{y}_2 = z\hat{\gamma}$, and resulting estimates of β_1 should have lower MSE than if data set 1 alone were used, even though data set 1 does not have information on x_2. Simulation results show that the two-sample estimator outperforms SS2SLS in bias and efficiency, with significant differences.

In Case 4, data set 1 has a larger set of instruments, but a smaller sample size, than data set 2. Information from data set 2 can still be used to improve first-stage estimates in a manner similar to case 2. Even if the instruments are uncorrelated, this can be done controlling for the sample correlation between the instruments caused by sampling error, or simply implicitly imposing that they are uncorrelated in generating the fitted values. The simulation presented uses the partial-out method (in contrast, Case 3 imposed $Cov(z, x_2) = 0$, possibly explaining its efficiency gain). The bias results mirror Case 2's, confirming the bias reduction of the TS2SLS estimator. However, TS2SLS is less efficient overall than SS2SLS.

Case 5 appears similar to case 2, but the role of the second data set is different: it has fewer observations, but a richer set of covariates. In this case, data set 2 becomes the "second-stage" data set. TS2SLS may be preferred for efficiency purposes (thanks to the inclusion of x_1) even if β_1 is identified and consistently estimated with data set 1 alone. Alternatively, if $\mathrm{E}(\varepsilon_1|z) \neq 0$ but $\mathrm{E}(\varepsilon_1|z, x_1) = 0$, estimates with data set 1 alone are inconsistent, but estimates using TS2SLS are consistent. In this scenario, the inclusion of data set 2 allows for an instrument which is only exogenous conditional on some vector of covariates x_1 to be validly used while still maximally exploiting the larger sample size of data set 1.Simulation results for this case are currently inconclusive.

Conclusions:

Provided two random samples from the same population, there are multiple scenarios of covariate availability in two data sets that have preferable estimates using the TS2SLS-type procedures presented here. The primary hypothesized improvements from data combination in the model presented are finite sample bias reduction and efficiency gain under certain conditions. The efficiency gain is only achieved if the reduction in overall sampling error achieved by the data combination exceeds the new sampling error induced by the "partialling out" procedure sometimes required to combine information on the basis of common covariates (or, in other words, the use of imputed regressors in the first stage). For example, Case 1, which did not have additional covariates beyond the endogenous regressor and the instrument, has an unequivocal efficiency gain, but the remaining cases are susceptible to lower overall efficiency. These results match theoretical expectations, but are only confirmable once the finite sample properties are analytically derived, and lacking that, once simulations are run with a wide range of parameter choices.

The chief limitations on these findings relate to the comparability of available samples. The common covariates used to link the data sets must measure the same thing, and may be measured with different levels of error. More critically, samples are rarely from the same population (and rarely truly random), and so optimal methods of sample reconstruction must be considered. Drawing on sample selection and quasi-experimental methods literature, different methods should be considered for augmenting auxiliary data in order to ensure the comparability of auxiliary and primary data (as measured by equivalence of sample moments). These include inverse probability weighting (Wooldridge, 2002), inverse probability tilting (Graham, Pinto, and Egel, 2008), imposing moment restrictions by weighting (Hellerstein and Imbens, 1999), and "entropy balancing" (Hainmueller, 2011).

There are several practical situations emulating the stylized cases in Table 1 that are amenable to TS2SLS, with the main practical constraint being data availability. The most likely candidate causal questions for this approach are those which are answered with instrumental variables that are universally available (e.g., birth date), are easily matchable from outside sources (e.g., IVs related to time and/or geography), or that relate to some independently known selection rule (i.e., regression discontinuity designs). Educational policy evaluations with a quasi-experimental approach are likely to benefit, as data sets often will have information on students' geographic regions, school districts, and even schools, which are easily matchable to policy variation affecting variables of causal interest.

Another broader example of an application lies in instrumental variables approaches to estimating the average return to schooling. Angrist and Krueger (1991) use compulsory school attendance laws and variation in age at school entry due to school start age policies to estimate a relationship between years of schooling and age at school entry, and subsequently use this exogenous variation in years of schooling to estimate the wage return to schooling. They use 2SLS to estimate a linear model with some controls using the 1970 Census, uniformly finding no significant difference between 2SLS and OLS estimates of the return to schooling. The combination of this Census data with a covariate-rich data set using the methods outlined here (i.e., Case 5), such as the Panel Study of Income Dynamics, can allow for more efficient estimates as more of the non-educational determinants of wages can be controlled for, even though they are not required for consistent estimation.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Angrist, J. D., & Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. doi:10.2307/2937954

Angrist, J. D., & Krueger, A. B. (1995). Split-Sample Instrumental Variables Estimates of the Return to Schooling. *Journal of Business & Economic Statistics*, *13*(2), 225–235. doi:10.2307/1392377

Angrist, J. D., & Krueger, A. B. (1992). The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples. *Journal of the American Statistical Association*, *87*(418), 328–336. doi:10.1080/01621459.1992.10475212

Arellano, M., & Meghir, C. (1992). Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets. *The Review of Economic Studies*, *59*(3), 537–559. doi:10.2307/2297863

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, *90*(430), 443–450. doi:10.2307/2291055

Graham, B. S., Pinto, C. C. de X., & Egel, D. (2011). Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST). *National Bureau of Economic Research Working Paper Series*, *No. 16928*. Retrieved from http://www.nber.org/papers/w16928

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, *20*(1), 25–46. doi:10.1093/pan/mpr025

Hellerstein, J. K., & Imbens, G. W. (1999). Imposing Moment Restrictions from Auxiliary Data by Weighting. *Review of Economics and Statistics*, *81*(1), 1–14. doi:10.1162/003465399557860

Inoue, A., & Solon, G. (2010). Two-Sample Instrumental Variables Estimators. *Review of Economics and Statistics*, *92*(3), 557–561. doi:10.1162/REST_a_00011

Klevmarken, N. A. (1982). On the Stability of Age-Earnings Profiles. *The Scandinavian Journal of Economics*, *84*(4), 531–554. doi:10.2307/3439516

Prokhorov, A., & Schmidt, P. (2009). GMM redundancy results for general missing data problems. *Journal of Econometrics*, *151*(1), 47–55. doi:10.1016/j.jeconom.2009.03.010

Ridder, G., & Moffitt, R. (2007). *The Econometrics of Data Combination* (Handbook of Econometrics). Elsevier. Retrieved from http://econpapers.repec.org/bookchap/eeeecochp/6b-75.htm

Wooldridge, J. (2002). Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, *1*(2), 117–139. doi:10.1007/s10258-002-0008-x

Appendix B. Tables and Figures

Not included in page count.

Table 1. Data Combination Situations

Case	Data set 1	Data set 2	Conditions
0	$\{y_1, z\}$	$\{y_2, z\ \}$	--
1	$\{y_1, y_2, z\}$	$\{y_2, z\}$	$N_1 < N_2$
2	$\{y_1, y_2, x_1, z\}$	$\{y_2, z\}$	$N_1 < N_2$
3	$\{y_1, y_2, x_1, z\}$	$\{y_2, x_2, z\}$	$Cov(z, x_2) = 0$
4	$\{y_1, y_2, z_1, z_2\}$	$\{y_2, z_1\}$	$N_1 < N_2$
5	$\{y_1, y_2, z\}$	$\{y_1, y_2, x_1, z\}$	$N_1 > N_2,$ $Cov(z, x_1) = 0$

Case 1 Simulation: Data Generating Process	
Data Set 1 Sample Size (N1)	200
Data Set 2 Sample Size (N2)	4,800
DGP	
	$z \sim N(0,1)$ $\varepsilon_2 \sim N(0,1)$ $u_1 \sim N(0,4)$ $u_2 \sim N(0,1)$ $\varepsilon_1 = 2\varepsilon_2 + u_1$ $x_1 = 2z + u_2$ $y_2 = z + 4\varepsilon_2$ $y_1 = y_2 + \varepsilon_2$

Case 1 Simulation Results (# Simulations = 200,000)		
Estimator	Mean	Standard Deviation
Full-sample First Stage F-Statistic	301.0804	35.7702
Small-sample First Stage Coefficient	0.999422	0.285607
Full-sample First Stage Coefficient	1.000047	0.057789
Full-sample OLS	1.470599	0.007218
Small-sample 2SLS ($\hat{\beta}_{SS,2SLS}$)	0.944031	5.584372
Hypothetical Full-sample 2SLS ($\hat{\beta}_{FS,2SLS}$)	0.998256	0.041193
Two-sample 2SLS ($\hat{\beta}_{TS2SLS}$)	1.002295	0.458146

Case 2 Simulation: Data Generating Process	
Data Set 1 Sample Size (N1)	3,000
Data Set 2 Sample Size (N2)	3,000
DGP	
	$z \sim N(0,1)$ $\varepsilon_2 \sim N(0,1)$ $u_1 \sim N(0,4)$ $u_2 \sim N(0,1)$ $\varepsilon_1 = 2\varepsilon_2 + u_1$ $x_1 = 2z + u_2$ $y_2 = z + 2\varepsilon_2$ y_1 $= y_2 + 10x_1 + 2\varepsilon_2$

Case 2 Simulation Results (# Simulations = 100,000)		
Estimator	Mean	Standard Deviation
Full-sample First Stage F-Statistic	2401.083	109.5684
Small-sample First Stage Coefficient	1.00034	0.224158
Full-sample First Stage Coefficient	0.999974	0.020418
Full-sample OLS	1.952388	0.010184
Small-sample 2SLS ($\hat{\beta}_{SS,2SLS}$)	0.938835	0.402948
Hypothetical Full-sample 2SLS ($\hat{\beta}_{FS,2SLS}$)	0.997849	0.064915
Two-sample 2SLS ($\hat{\beta}_{TS2SLS}$)	0.953917	0.409761

Case 3 Simulation: Data Generating Process	
Data Set 1 Sample Size (N1)	3,000
Data Set 2 Sample Size (N2)	3,000
DGP	
	$z \sim N(0,1)$ $\varepsilon_2 \sim N(0,1)$ $u_1 \sim N(0,4)$ $u_2 \sim N(0,1)$ $\varepsilon_1 = 2\varepsilon_2 + u_1$ $x_1 = 2z + u_2$ $x_2 \sim N(0,1)$ $y_2 = z + 10x_2 + 2\varepsilon_2$ $y_1 = y_2 + x_1 + \varepsilon_2$

Case 3 Simulation Results (# Simulations = 100,000)		
Estimator	Mean	Standard Deviation
Full-sample First Stage F-Statistic	37887.59	1414.267
Small-sample First Stage Coefficient	0.99928	0.416349
Full-sample First Stage Coefficient	0.999887	0.036519
Full-sample OLS	1.038353	0.005004
Small-sample 2SLS ($\hat{\beta}_{SS,2SLS}$)	0.88087	17.59361
Hypothetical Full-sample 2SLS ($\hat{\beta}_{FS,2SLS}$)	0.990412	4.778995
Two-sample 2SLS ($\hat{\beta}_{TS2SLS}$)	1.04026	9.535611

Case 4 Simulation: Data Generating Process	
Data Set 1 Sample Size (N1)	400
Data Set 2 Sample Size (N2)	9,600
DGP	
	$z_1 \sim N(0,1)$ $z_2 \sim N(0,1)$ $\varepsilon_2 \sim N(0,1)$ $u_1 \sim N(0,4)$ $u_2 \sim N(0,1)$ $\varepsilon_1 = 4\varepsilon_2 + u_1$ $x_1 \sim N(0,4)$ $y_2 = z_1 + z_2 + 16\varepsilon_2$ $y_1 = y_2 + \varepsilon_2$

Case 4 Simulation Results (# Simulations = 100,000)		
Estimator	Mean	Standard Deviation
Full-sample First Stage F-Statistic	38.3442	12.32802
Small-sample First Stage Coefficient	1.004556	0.803622
Full-sample First Stage Coefficient	0.999825	0.163579
Full-sample OLS	1.248064	0.001294
2SLS on Data Set 1 Alone	1.052799	0.35121
2SLS on Data Set 2 Alone	0.999892	0.033188
Hypothetical Full-sample 2SLS ($\hat{\beta}_{FS,2SLS}$)	1.0384	0.628719
Two-sample 2SLS ($\hat{\beta}_{TS2SLS}$)	38.3442	12.32802

Case 5a Simulation: Data Generating Process	
Data Set 1 Sample Size (N1)	400
Data Set 2 Sample Size (N2)	9,600
DGP	
	$z \sim N(0,1)$ $\varepsilon_2 \sim N(0,1)$ $u_1 \sim N(0,4)$ $u_2 \sim N(0,1)$ $\varepsilon_1 = 2\varepsilon_2 + u_1$ $x_1 \sim N(0,4)$ $y_2 = z + 6\varepsilon_2$ $y_1 = y_2 + 6x_1 + \varepsilon_2$

Case 5a Simulation Results (# Simulations = 200,000)		
Estimator	Mean	Standard Deviation
Full-sample First Stage F-Statistic	267.656	33.14328
Small-sample First Stage Coefficient	1.000627	0.301397
Full-sample First Stage Coefficient	0.999933	0.06124
Full-sample OLS	1.324331	0.003404
2SLS on Data Set 1 Alone	0.998517	0.126608
2SLS on Data Set 2 Alone	0.947522	2.713303
Hypothetical Full-sample 2SLS ($\hat{\beta}_{FS,2SLS}$)	0.998693	0.029129
Two-sample 2SLS ($\hat{\beta}_{TS2SLS}$)	1.004571	0.420684

Procedures

Case 1

1. Regress y_2 on z and get coefficient estimate $\hat{\gamma}$ in data set 2.
2. Generate fitted values $\hat{y}_2 = \hat{\gamma}z$ in data set 1
3. Use $\hat{y}_2$ as instruments for y_2 in a regression of y_1 on y_2.

Case 2

Here, the procedure is

1. Regress y_2 on z and get coefficient estimate $\hat{\gamma}$ in data set 2.

Then, in data set 1,

2. Generate residuals $\hat{e}_{y2} = y_2 - \hat{\gamma}z$
3. Regress x_1 on z and get residuals, $\widehat{e_z}$
4. Regress $\hat{e}_{y2}$ on $\widehat{e_z}$ and calculate fitted values $\tilde{y}_2$
5. Generate $\hat{y}_2 = \hat{e}_{y2} + \tilde{y}_2$
6. Use $\hat{y}_2$ as an instrument for y_2 in a regression of y_1 on y_2 and x_1

In a single sample, the $\hat{y}_2$ generated by this procedure is computationally identical to the fitted values from OLS of y_2 on z and x_1.

Case 3

1. Regress y_2 on z and x_2, and get coefficient estimate $\hat{\gamma}$, in data set 2.
2. Generate fitted values $\widehat{y_2} = \hat{\gamma}z$ in data set 1
3. Use $\hat{y}_2$ as instruments for y_2 in a regression of y_1 on y_2.

Case 4

1. Regress y_2 on z_1 and get coefficient estimate $\hat{\gamma}_1$ in data set 2.

Then, in data set 1,

2. Generate residuals $\hat{e}_{y2} = y_2 - \hat{\gamma}_1 z_1$
3. Regress z_2 on z_1 and get residuals, $\widehat{e_z}$
4. Regress $\hat{e}_{y2}$ on $\widehat{e_z}$ and calculate fitted values $\tilde{y}_2$
5. Generate $\hat{y}_2 = \hat{e}_{y2} + \tilde{y}_2$
6. Use $\hat{y}_2$ as an instrument for y_2 in a regression of y_1 on y_2 and x_1

Case 5

1. Regress y_2 on z and get coefficient estimate $\hat{\gamma}$ in data set 1.

Then, in data set 2,

2. Generate residuals $\hat{e}_{y2} = y_2 - \hat{\gamma}z$
3. Regress x_1 on z and get residuals, $\widehat{e_z}$

4. Regress $\hat{e}_{y2}$ on $\widehat{e_z}$ and calculate fitted values $\tilde{y}_2$
5. Generate $\hat{y}_2 = \hat{e}_{y2} + \tilde{y}_2$
6. Use $\hat{y}_2$ as an instrument for y_2 in a regression of y_1 on y_2 and x_1