Reassessment of classic case studies in labor economics with new instrument-free methods

Jan F. Kiviet*† and Sebastian Kripfganz‡

5 July 2020

JEL-Classifications: C10, C26, I26, J31

Keywords: *endogeneity robust least-squares inference, instrument validity tests, return to schooling, replication studies, sensitivity analysis.*

Abstract

It is demonstrated that even serious instrument invalidity will frequently not be detected by over-identification restriction tests, whereas instrumental variables based inference is shown to be seriously vulnerable regarding invalidity of instruments (even for mild invalidity, especially in combination with instrument weakness). Meanwhile, recently developed instrument-free methods provide valid confidence intervals for regression coefficients. These achieve set-identification through adopting credible ranges for endogeneity correlations. Besides, they provide more decisive evidence on invalidity of exclusion restrictions than Sargan-Hansen tests can. This all is illustrated for two classic case studies in labor economics focusing on the effect of education on earnings.

1. Introduction

Replication studies usually involve the application of established research methods to new data, in order to find out whether earlier empirical findings prove to have wider relevance. However, a replication study may also entail the application of new techniques to data that have earlier been analyzed by established state-of-the-art methods with the purpose to investigate whether the new techniques reinforce the credibility of the

*Amsterdam School of Economics, University of Amsterdam, PO Box 15867, 1001 NJ Amsterdam, The Netherlands (J.F.Kiviet@uva.nl).

†Department of Economics, University of Stellenbosch, South Africa.

‡University of Exeter Business School, Streatham Court, Rennes Drive, Exeter EX4 4PU, UK (S.Kripfganz@exeter.ac.uk).

original results or provide new insights. Here we apply recently developed instrument-free inference techniques to data earlier used in two classic case studies in labor economics that to date, according to current best practice, would still be analyzed by instrumental variables based techniques. From the results we conclude that, although both approaches are not without particular downsides, the kind of credibility that can be substantiated for findings obtained by the instrument-free approach are of a more comprehensive and solid nature than those at hand when using instruments.

Instrumental variables are variables which should be uncorrelated with the model error term. Exogenous regressors establish valid internal instruments. For models with some endogenous regressors as well, instrumental variable based inference requires at least as many external instrumental variables as there are endogenous regressors. These external instruments should validly be excluded from the regression model, implying that they are uncorrelated with the model disturbance term. Instrumental variables based techniques, already in use from early on in the 20th century, popular in particular in applied macroeconomics during some decades after the Second World War, gained firmer ground too in applied microeconomics, especially during the last decade before the turn of the century. These microeconometric studies stressed that the good reputation of causality studies, when based on well-designed controlled experiments, could be matched by observational data based studies too, provided these could be satisfactorily framed as so-called natural experiments. This inspired a revival of applications using the two-stage least-squares (TSLS) technique. However, these studies using instrumental variables did receive criticism for various good reasons; see, for instance, Bound, Jaeger and Baker (1995), Staiger and Stock (1997) and Rosenzweig and Wolpin (2000). Apart from potential invalidity (correlation between instrument and error) another ominous vulnerability concerns the possible weakness of external instruments. This occurs when the variation in endogenous regressors shows only little coherence with variation in the external instruments. Stock et al. (2002) argue that features that make it plausible for instruments to be exogenous can also make the instruments weak. However, due to further technical and methodological developments, TSLS and its generalization GMM (generalized method of moments) remained prominent and broadly respected tools in

modern applied econometric research. Of the serious weaknesses, as collected in the overview by Murray (2006), various have been addressed in the recent literature, see Andrews, Stock and Sun (2019).

These weaknesses of instrument-based analyses are in essence fourfold: (i) their standard normal asymptotic inference is highly inaccurate when using weak though valid external instruments (coefficient estimates are biased, they are non-normal in finite samples, and their standard deviation estimates lead to very poor size control of tests); (ii) although weak-instrument robust techniques improve the level control (provided the instruments are valid), they do yield confidence sets which are as a rule very wide or even unbounded; (iii) in the context of instrument-based inference, trustworthy statistical evidence on (in)validity of instruments can only be produced when a sufficient number of genuinely valid instruments is already available, so statistical evidence certifying validity of all employed instruments is an impossibility; and (iv) self-evidently, instrument-based inference will be seriously inaccurate if some instruments are in fact invalid, and can even be worse than inconsistent OLS-based inference.[1]

Addressing the problems (i) and (ii) is still receiving a lot of attention in econometric theory currently, but at the very best it will ultimately just yield appropriately size-controlled though very inefficient inference, which due to (iii) will always be built on insecure orthogonality assumptions. So, due to (iii) and (iv), putting trust in instrument-based inference will always be risky and often controversial.

The instrument-free approach illustrated here can provide some help to instrument-based inference to overcome problem (iii), which then might be beneficial to avoid problem (iv). More importantly, however, it enables to produce instrument-free inference on the specified regression model as such, which is therefore immune to the problems (i) through (iv) altogether. Needless to remark that this alternative approach goes with some particular problems of its own, as will be exposed below.

Earlier studies addressing problem (iv) did all stick to employing instrument-based techniques. Kraay (2012) used Bayesian methods allowing for a certain degree of instru-

[1] For (i) see Nelson and Startz (1990), Stock et al. (2002); for (ii) see Andrews and Stock (2007); for (iii) see Parente and Santos Silva (2012); and for (iv) see Kiviet and Niemczyk (2012).

ment invalidity. Nevo and Rosen (2012) derive set estimates under assumptions on the signs and relative magnitudes of the simultaneity and instrument invalidity. Conley et al. (2012) augment the model with the external instruments and make assumptions on their coefficients (which would be zero under correct exclusion). This allows frequentist or Bayesian methods to obtain inference allowing for instrument invalidity. These three approaches, though, are all still facing problems (i) and (ii), which can be circumvented by the instrument-free approach. Because any particular separate approach will be built on disjunct though unverifiable subjective assumptions, it seems wise for practitioners to adopt an eclectic attitude, in which findings from various alternative approaches are confronted with each other.

Inference on linear regression models with endogenous regressors, which does not use external instruments, has been developed in Kiviet (2020a, 2020b). It is based on bias-corrected least-squares estimation. As is well known, the least-squares estimator is inconsistent when one or more regressors are correlated with the error term. Obtaining nevertheless an assessment of this bias from this inconsistent estimator itself may at first sight appear quite eccentric. Therefore, this instrument-free least-squares based consistent estimator is addressed as kinky least-squares (KLS).

Correction for finite sample bias of regression coefficient estimators is usually only employed to consistent estimators; see, for instance, Kiviet and Phillips (1993) and MacKinnon and Smith (1998). The bias/inconsistency of least-squares in the present model is in fact a function of the vector of correlations between the regressors and the error term, to be indicated by ρ_{xu} below. Vector ρ_{xu} is generally unknown, and can only be estimated consistently on the basis of consistent residuals, which would require a consistent estimator of the coefficients. The latter can be obtained, of course, by exploiting valid external instruments, but this is a shaky source for achieving identification that KLS wants to avoid. Therefore, it pursues identification by making point or interval assumptions on the actual numerical value of ρ_{xu}. By adopting for each element of ρ_{xu}, either orthogonality of that regressor with respect to the error, or an interval regarding its possible nonorthogonality, set-identification is achieved, as defined in Bontemps and Magnac (2017), provided the specified zeros and intervals cover the true ρ_{xu}. It appears,

that the credibility one is willing to ascribe to either instrument-based or instrument-free inference will be determined unavoidably by subjective assessments of, either the claimed validity of the instruments, or the alleged reliability of the adopted numerical range of values for the degree of endogeneity. Here KLS seems to have a clear advantage, because its assumed set of correlation values does not have measure zero, as is the case for TSLS and GMM.

For a biased but consistent estimator the order of magnitude of a consistent estimator of its bias is of lower order in terms of the sample size than the estimator's distribution. In such cases, the leading term of the asymptotic variance of the bias-corrected estimator is equivalent to that of the asymptotic variance of the uncorrected estimator. For an inconsistent coefficient estimator, however, a consistent estimator of its bias is of such an order that the leading terms of the asymptotic variance of the uncorrected and the bias corrected coefficient estimators will be different. Although obtaining an expression for the consistent KLS estimator itself is quite straightforward, the derivation under the usual regularity conditions of its asymptotic variance, which is required for testing and confidence region construction, proves to be quite cumbersome. Simulation experiments in Kiviet (2020a,b) demonstrate, though, that in general the obtained asymptotic approximation to the actual distribution of the KLS estimator, unlike that of TSLS, is actually extremely accurate, even in very small samples.

In Section 2, we first highlight some basics about the instrument-free approach. Next, in Section 3, we compare the major hurdles affecting the approaches based on either exploiting instrumental variables or avoiding the use of instruments. Qualitative and quantitative illustrations of all this are based on a simulation study; its details are presented in Appendices (separately available as Supplementary material) A and B. Section 4 first analyzes some particular aspects of the endogeneity and instrumentation problems affecting the empirical analysis of earnings equations, with full technical details on that in Appendix C. Next, we contrast empirical results from the instrument-free and two instrument-based approaches, namely standard TSLS and the union of confidence intervals plausibly exogenous variant of TSLS suggested by Conley et al. (2012). Doing this for the two very well-known classic studies on the causal effect of education on

earnings by Angrist and Krueger (1991) and Card (1995), the readers are invited to form their own opinions and preferences. Finally, Section 5 concludes.

2. The essentials of KLS

Just for clarification, we provide here some more technical details on instrument-free inference. We shall present formulas for the KLS estimator and its variance estimator for the single coefficient β of a regression model for regressand y with just one endogenous regressor x and disturbances u, where for the identically and independently distributed observations $i = 1, ..., n$ we have

$$u_i \sim (0, \sigma_u^2),\ x_i \sim (0, \sigma_x^2),\ \text{with } E(x_i u_i) = \rho_{xu}\sigma_x\sigma_u. \tag{2.1}$$

Applying ordinary least-squares (OLS) yields the estimators

$$\hat{\beta}_{OLS} = \Sigma_{i=1}^n x_i y_i / \Sigma_{i=1}^n x_i^2,\ \hat{u}_i = y_i - \hat{\beta}_{OLS} x_i \text{ and } \hat{\sigma}_u^2 = n^{-1}\Sigma_{i=1}^n \hat{u}_i^2, \tag{2.2}$$

which are inconsistent for β, u_i and σ_u^2 when scalar $\rho_{xu} \neq 0$. KLS, which is consistent and asymptotically normally distributed when ρ_{xu} is known, is in this simple context defined by the estimators

$$\hat{\beta}_{KLS}(\rho_{xu}) = \hat{\beta}_{OLS} - \rho_{xu}\sqrt{\frac{\hat{\sigma}_u^2(\rho_{xu})}{n^{-1}\Sigma_{i=1}^n x_i^2}}, \tag{2.3}$$

$$\hat{\sigma}_u^2(\rho_{xu}) = \hat{\sigma}_u^2/(1 - \rho_{xu}^2), \tag{2.4}$$

$$\widehat{Var}\left(\hat{\beta}_{KLS}(\rho_{xu})\right) = \frac{4 + (\hat{\kappa}_x + \hat{\kappa}_u - 14)\rho_{xu}^2 - 2(\hat{\kappa}_u - 5)\rho_{xu}^4}{4(1 - \rho_{xu}^2)^2} \frac{\hat{\sigma}_u^2(\rho_{xu})}{\Sigma_{i=1}^n x_i^2}, \tag{2.5}$$

where $\hat{\kappa}_x$ and $\hat{\kappa}_u$ are the kurtosis estimators

$$\hat{\kappa}_x = \frac{n^{-1}\Sigma_{i=1}^n x_i^4}{(n^{-1}\Sigma_{i=1}^n x_i^2)^2},\ \hat{\kappa}_u = \frac{n^{-1}\Sigma_{i=1}^n (y_i - \hat{\beta}_{KLS} x_i)^4}{[n^{-1}\Sigma_{i=1}^n (y_i - \hat{\beta}_{KLS} x_i)^2]^2}.$$

So, the fourth moments of both x and u have an effect on the variance. If both happen to have kurtosis 3 then the true variance specializes to the familiar expression $\sigma_u^2/\Sigma_{i=1}^n x_i^2$, which is invariant regarding ρ_{xu}, and is consistently estimated by $\hat{\sigma}_u^2(\rho_{xu}^2)/\Sigma_{i=1}^n x_i^2$.

We conclude that, if ρ_{xu} were known indeed, producing inference on β in the form of tests and confidence regions would be easy. Testing the hypothesis $\beta = \beta_0$, where β_0 is

a known constant, against a one- or two-sided alternative requires confronting the test statistic

$$[\hat{\beta}_{KLS}(\rho_{xu}) - \beta_0]/[\widehat{Var}\left(\hat{\beta}_{KLS}(\rho_{xu})\right)]^{1/2} \tag{2.6}$$

with a standard normal critical value, or its square with one from $\chi^2(1)$. The endpoints of an $(1-\alpha)\times 100\%$ asymptotic confidence interval for β are given by

$$\hat{\beta}_{KLS}(\rho_{xu}) \pm \zeta_{1-\alpha/2}[\widehat{Var}\left(\hat{\beta}_{KLS}(\rho_{xu})\right)]^{1/2}, \tag{2.7}$$

where ζ_p is the p^{th} quantile of the standard normal distribution. Such infeasible (because ρ_{xu} is generally unknown) KLS inference on β has proved to possess highly desirable properties, because $\hat{\beta}_{KLS}(\rho_{xu})$ is virtually unbiased in finite samples of typical cross-section models. Moreover, its variance is as a rule smaller than that of instrumental variables based estimators, especially when based on weak instruments. When instruments are invalid, which renders TSLS inconsistent, consistent KLS is certainly much more attractive.

However, KLS is only feasible when ρ_{xu} is known, whereas in practice ρ_{xu} is generally unknown. Therefore, KLS inference should be produced over a range of chosen realistic values r_{xu}. This is easily done in practice, but indicating which values are unreasonable, and which are reasonable, requires subject matter knowledge not always available, and if available not always trustworthy. This is comparable to the situation for instrument-based analysis, where statements on the assumed validity of instruments may lack persuasiveness. A crucial difference is, however, that it is not at all straightforward to investigate the sensitivity of TSLS to a certain degree of invalidity of instruments, whereas for KLS the sensitivity of inference on regression coefficients regarding ρ_{xu} is an intrinsic part of the instrument-free approach, as will be exposed below.

When the model contains exogenous regressors too, or if the model has more than one endogenous regressor, more general formulas than those given above apply, see Kiviet (2020b). Also models with dependent time-series observations can be handled. The more general framework can also easily be used to test whether particular regressors seem omitted from the model, which enables to test the exclusion restrictions which are so crucial for an instrument-based analysis. KLS-based confidence sets for the coefficients of

excluded regressors can be used as input for the Conley et al. (2012) plausibly exogenous techniques. Moreover, KLS easily allows to implement tests for misspecification, such as for heteroskedasticity, structural change, serial correlation (just relevant in a time-series context) or RESET tests for improper functional form, all without having to adopt instrumental variables.

3. The impediments of instrument-based and instrument-free inference

When regression models contain endogenous explanatory variables, identification of the regression coefficients is troublesome. A consequence may be that estimators may systematically deviate from the aimed at true parameter value. The standard approach is to find so-called external instrumental variables which, unlike the endogenous regressors, should be uncorrelated with the disturbance term, whereas at the same time their correlation with the endogenous regressors should not be weak. So, the causal effect of the external instruments on the dependent variable should not be direct, but exclusively and significantly indirect via its effect on the endogenous explanatory variables.

For consistent estimation one needs at least as many external instruments as the model contains endogenous regressors. These external instruments are uncorrelated with the disturbance term indeed when they are validly excluded from the regression equation under study. Testing whether they are correctly excluded is seriously problematic, because this requires consistent estimation of the model augmented with these candidate external instruments. As it seems, such testing would only be possible if further valid external instruments were available, which could only be tested themselves with even more validly excluded instruments, and so on. Therefore, exclusion restrictions tests are just viable for the so-called over-identification restrictions, being the subset of the putative external instruments not required for the consistent estimation of the (now just-identified) extended model. Hence, unavoidably within this context, the valid exclusion of a number of external instruments equal to the number of endogenous regressors has simply to be adopted on the basis of other information than statistical testing.

For standard linear models, the over-identification restrictions test is the Sargan test.

For its correspondences with testing exclusion restrictions, see Kiviet (2017). Because of the difficulties mentioned above, this test is not consistent for validity of all external instruments: even in infinitely large samples it may reject with probability smaller than one when some of the external instruments are in fact invalid. This seriously undermines the credibility of the test to properly provide information on instrument validity: an insignificant test result can be due both to validity and to invalidity of all or some of the exclusion restrictions. In addition to this problem, one finds in the literature that the test is often blamed for over-rejection in finite samples. Hence, apparently a significant outcome of the test can also be due both to valid and to invalid over-identification restrictions. Therefore, some authors advise practitioners to use the test at a very low nominal significance level; see Hansen (2019, p.434).

On the other hand, however, one may also argue in favor of testing at a very high nominal significance level, because of the following. An insignificant value of the test is used in practice to accept the null hypothesis of validity of all external instruments. However, statistical tests are designed to support the decision to reject (or not) a null hypothesis. This is at odds with ever deciding to wholeheartedly accepting it. So, in this curious case, our primary worry should be to fail to reject invalid instruments (commit a type II error) and not so much to limit type I errors (wrongly rejecting valid instruments). Therefore, we are in fact inclined to advise to be extremely prudent, by deciding to accept instrument validity and the corresponding TSLS results only when the p-value of the Sargan test is pretty high; perhaps only when it is larger than 50%, instead of the habitual 5%! However, to decide on a really sensible nominal level α of the over-identification restrictions test, one should know what the price is of using invalid instruments.

It is quite remarkable that despite the frequent use of TSLS (and its generalization GMM) there is very little concrete information in the literature on the actual performance of the Sargan test (and its generalization the Hansen test) and the consequences for inference on model coefficients of type II errors of these tests. For the special context of dynamic panel data model estimation by GMM some simulation studies are available, but for a standard TSLS application on cross-section data we were not able to

find useful well-founded guidelines. Therefore, we ran a small-scale simulation study in order to develop some useful intuition regarding the (lack of) qualities of the Sargan test. We examined its successes and failures to (in)correctly approve or reject external instruments as being valid under different situations regarding regressor endogeneity, instrument strength and degree of instrument (in)validity. Moreover, under these situations, and for different degrees of over-identification, we examined the distribution of the TSLS coefficient estimator, and compared it with KLS and OLS. All details on these experiments can be found in Appendices A and B; in this section we just report the major findings.

We simulated typical cross-section data for a relationship with just one endogenous regressor and an intercept, whereas there are two candidate external instruments. We avoided extreme cases of seriously weak or exceptionally invalid instruments. Samples of size 50, 250 and 2500 have been analyzed. We found that the Sargan test, when using both instruments and testing the single over-identification restriction, has actual type I error probabilities very close to the chosen nominal significance level α, for $0.05 \leq \alpha \leq 0.5$, when both instruments are valid indeed. As expected, when one or both instruments are invalid, the rejection frequency not only increases with α, but with the sample size too. However, the rejection frequency is barely larger than α for particular combinations of invalidity and strength of the instruments, even when the sample size is really large, especially when the more invalid instrument is relatively strong. Nevertheless, using $\alpha = 0.05$ may result in a rejection probability above 0.8 for particular correlation combinations when at least one of the instruments is invalid, especially when there is one valid and relatively strong instrument. Though, scrutinizing the detailed results in Appendix A, one should realize that the Sargan test is not a trustful guide. In Appendix B we show that in certain cases of serious instrument invalidity, where the two instruments have a similar ratio between their correlations regarding degree of invalidity and strength, the rejection probability of the Sargan test will always be close to the chosen significance level and thus lacks power.

In the simulations, we also investigated the effects of the various correlations and the sample size on the median and interquartile range of the slope estimator when just

one or both instruments are being used, and compared these with OLS and KLS. Of course, OLS is unbiased only when the regressor is exogenous, whereas its bias sharply increases for soaring endogeneity. In this simple static model, unfeasible KLS (which uses full knowledge of the actual endogeneity) is found to be median unbiased, and so are instrument-based estimators when the employed instruments are valid. Since the validity of instruments is in fact equally untraceable as the actual value of the endogeneity correlation, consistent instrument-based estimators are actually unfeasible as well. Of course, when instruments are invalid, the instrument-based estimators will be biased, but we find that this bias (unlike for OLS) is largely invariant regarding the degree of endogeneity. The interquartile range of KLS proves always much more attractive than that of instrument-based estimators, especially when weak and/or invalid instruments are being used. Also for the hazardous cases, where the Sargan test will reject invalid instruments with a disappointing probability very close to the significance level, the bias of instrument-based estimators is serious and the interquartile range much larger than for KLS.

So, undeniably, it may often happen that TSLS results will not be disapproved, because the Sargan test produces a pretty large p-value, although the instruments are actually invalid and generate inference of poor quality. Likewise, however, the accuracy of an assessment of the degree of endogeneity may be poor, so that feasible KLS may be seriously biased as well. Therefore, we also simulated the median (which proved to be invariant with respect to the sample size) and the interquartile range of feasible KLS. The results made us conclude that the vulnerability of KLS to moderate errors, although substantial, seems more limited than that of TSLS when using mildly invalid instruments. Moreover, KLS has an additional advantage: Whereas it is not self-evident to examine in practice the sensitivity of TSLS with respect to varying degrees of invalidity of the external instruments, the implementation of KLS which we shall use in the applications below incorporates by default an insightful analysis of its sensitivity regarding the actual degree of endogeneity.

4. Applications to classic studies on the return to schooling

The upsurge in the use of instrumental variables techniques in microeconomics in the 1990s was triggered in particular by novel studies in labor and especially in education economics, see the overviews in Card (1999, 2001) and Angrist and Krueger (2001). Below we re-analyze the original data sets from two very influential papers, namely Angrist and Krueger (1991) and Card (1995), where TSLS has been employed. We confront their results with KLS findings on the validity of the adopted exclusion restrictions. In addition, we will also compare the major inferences on the coefficients of primary interest as obtained from the instrument-free approach with those from instrument-based approaches, namely standard TSLS and one of its modifications as suggested by Conley et al. (2012). Moreover, we will employ instrument-free misspecification tests. These are found to detect model failures which previously remained unnoticed. We limit ourselves to just a few of the empirical results from the original published articles. Our selection here is strongly influenced by the particular IV/TSLS illustrations from these articles presented in Hansen (2019). Before we address the two classic studies, we first give in the next subsection some general background to the particular type of endogeneity and orthogonality assumptions made in this literature. Because these assumptions are usually left implicit, and do not always seem well understood, we produce a formal derivation of these in Appendix C.

4.1. Endogeneity and instrumentation when regressors have been omitted

The applications to be considered are characterized by the following. The relationships under study have the form of a linear (in the coefficients) regression model for which the regressors fall into three distinct categories. We denote this demeaned model therefore as

$$y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon, \tag{4.1}$$

where the X_j are $n \times K_j$ matrices and y and ε are $n \times 1$ vectors with $E(\varepsilon \mid X_1, X_2, X_3) = 0$. The $K_3 > 0$ variables X_3 are unavailable. Thus, the model to be estimated has regressors $X = (X_1, X_2)$ only. We assume $rank(X) = K_1 + K_2 > 0$. The distinction between the

regressors X_1 and X_2 is that vector β_1 contains the coefficients of primary interest, for which we are keen to find a consistent estimator.

Because the regressors X_3 are unavailable, the model has – next to the disturbance ε – an unknown individual effect represented by the $n \times 1$ component $\gamma = X_3\beta_3$. For this we suppose

$$\gamma = X_3\beta_3 \neq 0, \text{ where } E(\gamma \mid X) = X_1\phi_1 + X_2\phi_2. \tag{4.2}$$

Hence, we allow that X_3 is associated with X_1 and X_2. Substitution of (4.2) into (4.1) and defining $\eta = \gamma - E(\gamma \mid X)$, with $E(\eta \mid X) = 0$, gives

$$y = X_1(\beta_1 + \phi_1) + X_2(\beta_2 + \phi_1) + (\eta + \varepsilon), \text{ with } E(\eta + \varepsilon \mid X) = 0.$$

Thus, regressing y on X will yield least-squares coefficients that are consistent for $\beta_1 + \phi_1$ and $\beta_2 + \phi_2$, which represent the sum of the direct and, if any, the indirect effects (via X_3) of X_1 and X_2 on y. By using TSLS, though, it is possible to obtain under particular conditions consistent estimators for the direct effect β_1 of the regressors X_1.

Before we produce these particular conditions, we will first indicate some of the links of model (4.1) with the applications to follow. In these, the dependent variable y contains observations on the log of wage of individuals. Regressor X_1 just contains the explanatory variable schooling in years. Assessing its coefficient β_1 is the major goal of the analysis. Regressors X_2 concern control variables, such as gender, age, race, residence and time effects. Unobserved component γ represents the effects on wage of "ability", which is based on X_3 regressors like: having special skills, and notions expressing appearance, upbringing and charm. In model (4.1), which is assumed to contain all major explanatories, all variables in X_1, X_2 and X_3 are assumed to jointly cause y. They are exogenous, because no immediate feedbacks from y into any of these seem realistic. Some of the variables in X_1, X_2 and X_3 will be mutually related. Variables in X_2 can even be causal for X_1 (older generations may be less educated), and variables from X_3 will have an effect on X_1 too, because family background will be one of the determinants of the duration of education. So, X_3 will have direct effects on y, expressed by β_3, but also indirect effects through X_1, if $\phi_1 \neq 0$. However, neither schooling nor ability will be causes for the much more autonomous variables gender, age and race,

whereas these variables from X_2 are likely to have, next to a direct effect on y, also effects on X_1 and some of the variables in X_3, as already mentioned.

In Appendix C we derive the conditions under which regressing y on $X = (X_1, X_2)$, while using an instrumental variable matrix $Z = (Z_1, X_2)$, where Z_1 contains at least K_1 external instruments, will result in consistent estimation of the coefficients β_1 and $\beta_2^* = \beta_2 + \phi_2$, the sum of the direct effects β_2 from X_2 on y and any indirect effects from X_2 on y via unobserved component γ. These conditions, formulated in more technical terms in Appendix C, are as follows. Variables Z_1 should have correctly been omitted from the full model (4.1). This means that Z_1 has no direct causal effect on y. However, unless $\beta_1 = 0$, Z_1 should have an indirect effect on y via its association with X_1. If this association is substantial, this avoids the problem of weak instruments. Moreover, apart from any association with the variables in X_2 that the variables in Z_1 and X_3 may have, these two sets of variables (so after netting out their association with X_2) should be uncorrelated. Otherwise, TSLS will be inconsistent for the estimation of β_1.

Hence, X_1 should contain the regressors from X for which one wants to assess their direct causal effect on y. In the model that omits regressors X_3, an individual variable from X_1 becomes endogenous, if –after netting out any association this variable or component $\gamma = X_3\beta_3$ may have with the variables in X_2– they are still correlated. Note that by allocating more regressors from X_2 to X_1 (so aiming to estimate the direct causal effect of more explanatories of y) requires an extension of matrix Z_1, with the associated extra requirements regarding the uncorrelatedness of all its columns with γ (after netting out their association with the fewer variables in X_2).

When applying KLS, instead of finding external instruments Z_1, an assessment of the degree of endogeneity of the variables X_1 is required. When interpreting the coefficients of X_2 as an amalgam of direct and indirect effects, their correlation with the errors can safely be assumed to be zero.

4.2. Angrist and Krueger (1991)

In this article, referred to as AK below, a novel source for identification of the coefficient on years of education in a wage equation has been suggested, namely quarter of birth.

Because laws on compulsory school attendance differ by state in the US, there is a very moderate but distinctive source of variation in years of schooling due to quarter of birth. Assuming that quarter of birth has no direct effect on earnings, quarter of birth dummies would establish valid external instruments. An extremely large sample ($n = 329,509$) on individuals is available (1980 census: men born in 1930-1939). Most equations estimated by AK use very many (up to 180) instruments, which are all dummies and (subsets of) interactions of dummies for quarter of birth, year of birth and state of birth. However, many of these instruments turn out to be very weak, see Bound et al. (1995) and Staiger and Stock (1997). Therefore, in his illustrations explaining dependent variable *logwage*, Hansen (2019, p.459) decides to use just the 3 dummy external instruments constructed from the quarter of birth, which jointly seem sufficiently strong. Due to the omission of variables representing the effects of ability, intelligence, talent and the like, just the regressor education in years (*edu*) is supposed to be endogenous, whereas 20 further dummy controls are treated as exogenous, namely: race *(black)*, urban *(smsa)*, *married*, nine distinct year-of-birth dummies and eight particular region-of-residence dummies. The TSLS estimates are presented in Hansen's equation (12.96), whereas his equation (12.95) gives the corresponding reduced form equation for *edu*. It yields an F-test of 31 on the joint significance of the three quarter of birth (*qob*) dummies, so they seem sufficiently strong indeed.

We find the same TSLS results. Further calculations yield an estimate of ρ_1, the correlation of variable *edu* with the TSLS residuals, of -0.18 (using all 3 instruments), and of -0.47, -0.53 and -0.08, when using just one of the instruments *qob_2*, *qob_3*, and *qob_4*, respectively. Note, though, that in fact a positive ρ_1 is expected, because years of education is supposed to be positively correlated with the omitted explanatory component ability/skills. Of course, these estimates of ρ_1 are random, so may in fact not be significantly negative. Moreover, they are consistent for the correlation between *edu* and the unobserved model components (disturbance plus ability/skills effect) only, if no further relevant explanatory components of *logwage* have been omitted. Any further omitted regressors that are correlated with the external instruments would ruin easy interpretation of the present TSLS findings. As yet, the obtained negative ρ_1 estimates

provoke serious doubts about the current TSLS findings.

[Figure 4.2.1 here]

On this particular specification (12.96), that Hansen used in his textbook, and which has also been examined in Conley et al. (2012, p.269), we present KLS results[2] in four panels with graphs in Figure 4.2.1. The graph in the top-left panel shows asymptotic 95% confidence sets for the coefficient of *edu* for a wide range of adopted ρ_1 values. The KLS intervals vary substantially though rather systematically with the value of ρ_1. They are in fact so narrow, that they appear as one line in the graph. The graph also shows the much wider TSLS interval, which is invariant regarding ρ_1. We note that the KLS findings are in line with those obtained by TSLS if ρ_1 were mildly negative indeed, but are in sharp contrast for positive ρ_1 values. Both the KLS and TSLS intervals are based on the assumption that all regressors apart from *edu* are exogenous, which means that we should interpret their coefficients (not presented in this figure) as representing both the direct effects of these regressors and their indirect effects via the omitted regressors. The depicted inference on the coefficient value of *edu* represents just its direct effect, provided the endogeneity of *edu,* incurred due to the omission of explanatories that are correlated with *edu*, has adequately been accommodated. This requires for TSLS validity of the instruments, and for KLS focussing on a preferably narrow interval within (-1, 1), which should include the true value of ρ_1. By taking the union of all asymptotic 95% KLS confidence intervals for, for instance, $0 \leq \rho_1 \leq 0.2$, we may conclude that with an asymptotic confidence coefficient exceeding 95% the direct effect of *edu* is in the interval [0.021,0.064] if $\rho_1 \in [0, 0.2]$ indeed. By taking the union of one-sided asymptotic 97.5% confidence intervals, we may also conclude that, with an asymptotic confidence coefficient exceeding 97.5%, the direct effect of *edu* is positive provided $\rho_1 \leq 0.3$, or is smaller than 0.06 if $\rho_1 \geq 0$.

[2]The results for empirical data have all been obtained by Stata, employing for KLS the kinkyreg Stata program, soon to be contributed to the Stata community, and documented and illustrated in Kripfganz and Kiviet (2020).

The top-right panel of Figure 4.2.1 presents p-values of single and joint exclusion restrictions tests on the three external instruments over a wide range of postulated ρ_1 values. For the single tests, p-values of one arise for estimates $\hat{\rho}_1$ obtained by (just-identified) IV estimation. In Kiviet (2020b) it has been demonstrated that this will always happen, and does not carry any new information on possible validity of the instrument as such. It simply expresses that IV, which adopts one particular exclusion restriction, will yield residuals which have a correlation with the endogenous regressor for which KLS will support this exclusion restriction. In line with that, it is no surprise that the maximum p-value for the joint exclusion restrictions test is found for the value -0.18 of the TSLS estimate of ρ_1. What these graphs do portray is that, in case ρ_1 is actually positive, the obtained p-values provide evidence that feed serious doubt on the validity of the exclusion restrictions. Only when one has a priori reasons to believe that ρ_1 is negative, these graphs provide some support for validity of the instruments. And, vice versa, if one has a priori reasons to believe that the instruments must be valid, then the graphs disclose information that the true value of ρ_1 seems mildly negative.

If one finds it hard to believe that quarter of birth really has a direct effect on wage, additional to any indirect effect via education, invalidity of these external instruments does not seem due to wrongly excluding them as such, but to confronting them with the biased residuals of a misspecified relationship. Then it seems most likely that there are further omitted regressors, and that apparently the quarter of birth dummies happen to be correlated with these, which renders them invalid instruments anyhow. The bottom two panels of Figure 4.2.1 demonstrate that model specification (12.96), just allowing for endogeneity of *edu*, fails the various misspecification tests regarding heteroskedasticity[3] and RESET[4] for any value of ρ_1. Hence, no valid inferences on β_1 can be drawn from the top-left panel; so, a respecification of the model is called for.

[Figure 4.2.2 here]

[3]Here the joint significance is tested of the slopes in auxiliary regressions of the squared KLS residuals on an intercept and particular sets of regressors. In these sets X_2 refers to the exogenous regressors in the model, Z_1 to the external instruments, and X_1^{adj} to the estimated exogenous component of X_1.

[4]Here the joint significance is tested of the additional regressors, when the model is augmented by $\hat{y}_i^2, ..., \hat{y}_i^d$, where the integer order is $d \geq = 2$ and $\hat{y}_i$ is the estimated exogenous component of y_i.

We choose to augment the model with the three variables for which we found evidence that (when $\rho_1 > 0$, as theory suggests) their exclusion is rejected. Hence, we shall estimate now the model

$$y = X_1\beta_1 + X_2\beta_2 + Z_1\delta + \varepsilon, \tag{4.3}$$

where X_1 still has just one column (*edu*), and we use as controls both the former X_2 and the three quarter-of-birth dummies collected in Z_1. Only X_1 is treated as endogenous. Apart from analyzing this model by KLS, we will also apply the instrument-based union of confidence intervals technique suggested by Conley et al. (2012) and programmed in the *plausexog* Stata command by Clarke and Matta (2018). For that we have to adopt assumptions on the possible values of the three elements of vector δ. We can obtain empirical support for such assumptions by KLS as follows. The first three panels of Figure 4.2.2 present KLS inference on the coefficients of the *qob* dummies in model (4.3). Supposing that $0 \leq \rho_1 \leq 0.4$, we can choose for these intervals [-0.05, 0.01], [0.0, 0.02] and [-0.05, 0.02] respectively. The right-hand panel in the second row produces confidence intervals for the coefficient of *edu* obtained by applying KLS to (4.3) and also by using the plausibly exogenous approach. Note that the latter interval is invariant regarding ρ_1, overlaps the TSLS interval presented in Figure 4.2.1 because the chosen intervals do not exclude $\delta = 0$, has asymptotic confidence coefficient exceeding 95%, and is so wide that it is of little practical use.[5] It does not exclude that an extra year of schooling increases wage by either an outrageous 50% or even reduces it by not less than 20%. Hence, it seems that more efficient though endogeneity robustified inference can be obtained by bounding the endogeneity correlation of an endogenous regressor and apply KLS, than by bounding the degree of violation of exclusion restrictions to a realistic degree.

The bottom row of panels in Figure 4.2.2 shows that instrument-free heteroskedasticity and RESET tests yield very low p-values, irrespective of the chosen value for ρ_1. So, formally, these findings reject specification (4.3) very strongly. On the other hand, given

[5] Using the same data Conley et al. (2012) produce similar results, where all three *qob* coefficients are supposed to have a value in $[-\phi, \phi]$ for $\phi \in [0, 0.02]$. Focussing on $\phi = 0.01$ they also conclude that the data are essentilly uninformative about the returns to schooling.

the extraordinarily large size of the sample, one may expect that any specification of this relationship will fail, if it uses just a few dozen parameters. Nevertheless, we suppose that the model does require a serious respecification, as already had been suggested a long time ago in Bound et al. (1995), Bound and Jaeger (2000) and many other studies. We found that augmenting the model just by the controls age in years and its square, as suggested by Bound and Jaeger (1991), does not improve the situation. Hence, it seems that the set of explanatory variables included in this classic data set should be extended by further relevant explanatories, which is beyond the purpose of our primarily methodological replication study.

4.3. Card (1995)

Card examines the individual wage equation, too. His analysis is based on US survey data from 1976 and involves 3010 young men. Again, the coefficient of primary interest is the effect of years of education (*educ*) on the log of individual wage (*lwage*). Further covariates are experience (*exper*) and its square, and one ethnic (*black*) and two demographic dummy variables, namely *south* and *urb* (urbanization). As in AK, the additional effect of skills/ability (for which no data are available) are necessarily omitted, whereas these are supposed to be positively correlated with years of education. So again, this regressor should be positively correlated with the error term, and the education effect as obtained from OLS estimation should be positively biased. As an instrument, Card uses a dummy variable *college* indicating whether there is a college in the county where the young man concerned lives. A college in the proximity is supposed to have a direct positive effect on years of schooling, but no direct effect on wage. Hence, if this proximity variable is a valid instrument indeed, one would expect the IV estimate of the education coefficient to be smaller than the OLS coefficient. However, as in the foregoing subsection, it is not. One obtains (not robustifying the standard errors) for OLS 0.0740 (0.0035) and for IV 0.1323 (0.0492), whereas the correlation between the IV residuals and the education variable is -0.21. Note that the IV coefficient estimate is almost twice that of OLS, whereas its standard error (given in parentheses) is about 14 times as large as that of OLS.

The question is again: How to make sense of this? One possibility would be to simply blame weakness of the instrument for these estimation problems and contradictions, but there are other options too. A valid point raised by Card is that there may be other causes of endogeneity here than just omitted variables, such as measurement errors in years of education, which could explain a negative endogeneity correlation. Moreover, if education is endogenous, so will experience be, and also its square, because experience is constructed simply by subtracting education + 6 from age. Hence, we should allow for at least three endogenous regressors, and could use age and its square and the college dummy as instruments to achieve identification. Note, though, that a constant correlation between education and the error term implies exactly the opposite correlation between experience and the error term, and will have peculiar consequences for the endogeneity correlation of experience squared. This will complicate a KLS analysis. In the end, though, Card concludes that experience squared is in fact insignificant.

As applying KLS to models with more that one endogenous regressor –although possible– is a bit cumbersome too, we will choose a different road. Because

$$exper = age - educ - 6, \tag{4.4}$$

the model

$$lwage = \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + ... + u \tag{4.5}$$

implies

$$\begin{aligned} lwage &= \beta_1 educ + \beta_2(age - educ - 6) + \beta_3(age - educ - 6)^2 + ... + u \\ &= (\beta_1 - \beta_2 + 12\beta_3)educ + (\beta_2 + 12\beta_3)age + \beta_3 age^2 - 2\beta_3 age \times educ + ... + u \\ &= \theta_1 educ + \theta_2 age + \theta_3 age^2 + \theta_4 age \times educ + ... + u. \end{aligned} \tag{4.6}$$

If $\beta_3 = 0$, then $\theta_3 = \theta_4 = 0$ and $\beta_1 = \theta_1 + \theta_2$. Hence, inference on β_1 of (4.5) can in this case also be obtained from simply replacing the regressor experience by age and analyzing the sum of the education and age coefficients in (4.6). We will do so in the results below, but also keep initially age squared in the regression (but not the endogenous interaction of age and educ) and use, self-evidently, age and its square as internal instruments. Then

the model contains one endogenous regressor, has six exogenous regressors, and uses one external instrument.

Just identified TSLS for this restricted version of (4.6) yields (omitting presentation of the intercept)

$$\begin{array}{cccccc} lwage = 0.094educ + 0.082age - 0.073age^2/100 - 0.101black - 0.099south + 0.108urb \\ (0.050) \quad (0.071) \quad (0.124) \quad (0.075) \quad (0.030) \quad (0.050) \\ [0.049] \quad [0.070] \quad [0.123] \quad [0.073] \quad [0.030] \quad [0.050] \end{array}$$

Below the coefficients, we first present the usual standard error estimates in parentheses and in the next line their (hardly affected) robustified (regarding heteroskedasticity) versions in square brackets. In the reduced form equation for education, the regressor college has an F-value of 10.5, so although not distinctly weak the single instrument is certainly not strong either. Using, as Card does, two separate dummy instruments, by splitting *college* into presence in the county of *public* and *private* colleges, yields

$$\begin{array}{cccccc} lwage = 0.121educ + 0.054age - 0.023age^2/100 - 0.060black - 0.085south + 0.082urb \\ (0.038) \quad (0.065) \quad (0.114) \quad (0.059) \quad (0.026) \quad (0.040) \\ [0.038] \quad [0.065] \quad [0.113] \quad [0.058] \quad [0.027] \quad [0.040] \end{array}$$

Now, the relevant F-test is 10.2, so it has not improved. The Sargan test has p-value 0.46. Both being college dummies, it seems very unlikely that one of these instruments would be valid and the other invalid. From the simulations we learned that this p-value for the Sargan test may just as well mean that both instruments are valid or both invalid, which is not very reassuring.

Next, leaving out the insignificant age squared regressor gives

$$\begin{array}{cccccc} lwage = 0.121educ + 0.041age - 0.060black - 0.085south + 0.082urb \\ (0.038) \quad (0.003) \quad (0.060) \quad (0.026) \quad (0.041) \\ [0.038] \quad [0.003] \quad [0.058] \quad [0.027] \quad [0.040] \end{array} \tag{4.7}$$

This goes with an F-value of 10.2 and a Sargan p-value of 0.54, whereas the sum of the coefficients of education and age is estimated to be 0.162 with standard error 0.039,

which conforms extremely closely to the result obtained by Card when allowing for three endogenous regressors. Next, we shall examine what KLS yields for model (4.7).

[Figure 4.3.1 here]

Figure 4.3.1 contains four panels with graphs. The top-left panel shows what we can say about the magnitude of the direct effect θ_1 of education on log wage, assuming model specification (4.7) to be adequate, while we pretend to know the true value of ρ_1. KLS shows that this effect is positive, provided $\rho_1 < 0.18$, whereas for $\rho_1 > 0$ the effect is smaller than 0.04, which is substantially smaller than suggested by TSLS. However, despite the reassuring Sargan test, the top-right panel of the figure casts serious doubts on the validity of the instruments *public* and *private*. The graphs on the exclusion restrictions tests show the same pattern as in the foregoing subsection. On the basis of IV analysis, when just using the *public* college dummy as an instrument, residuals are obtained which have correlation -0.48 with the endogenous regressor *educ*ation, and when using just the *private* college dummy the estimate of ρ_1 is -0.66. When using both instruments in TSLS it is -0.47. Again, the KLS exclusion restrictions tests have very high p-values at these specific correlations. However, if ρ_1 is positive, as initially expected, then the exclusion restrictions are very strongly rejected. So, the initial conjecture that model (4.7) is well specified, apart from lacking the explanatory variable ability/skills, whereas proximity of a college has a positive effect on education but no direct effect on wage, is strongly knocked down. The panels in the bottom row of Figure 4.3.1 provide a mixed picture regarding the econometric checks on the adequacy of the specification. If for some yet not understood reason ρ_1 is negative indeed, then we find that the model could be affected by heteroskedasticity.

An obvious explanation for the invalidity of the instruments could be that families which endow their children with favorable abilities and skills do also actively choose to live near a college. Then the college variables would be invalid instruments and KLS would correctly detect that these have been wrongly excluded from specification (4.7). In a KLS analysis we can easily augment this model by these two extra controls, and

estimate this "underidentified" model without adopting any new external instruments. Just for illustrative purposes we present here KLS estimates for this augmented model while adopting the alternative identification assumption $\rho_1 = 0.1$. This mild endogeneity of *educ* yields

$$\begin{aligned} lwage = {} & 0.018educ + 0.039age - 0.212black - 0.129south + 0.171urb \\ & \quad (0.003) \qquad (0.002) \qquad (0.018) \qquad (0.016) \qquad (0.017) \\ & + 0.041public - 0.012private \\ & \quad (0.017) \qquad (0.021) \end{aligned} \tag{4.8}$$

We note that the former instruments are not both individually significant when $\rho_1 = 0.1$. Comparing with (4.7) KLS yields much smaller standard errors and substantially different coefficient values. More detailed results on model specification (4.8) for arbitrary ρ_1 can be found in Figures 4.3.2 and 4.3.3.

[Figure 4.3.2 here]

From the top row of Figure 4.3.2 one may infer that, supposing $0.0 \leq \rho_1 \leq 0.4$, the coefficients of the earlier excluded variables *public* and *private* may have values in the intervals [0.0, 0.1] and [-0.05, 0.04] respectively. In the left graph of the next row of Figure 4.3.2 these intervals have been used to apply the Conley et al. method to obtain a conservative asymptotic 95% confidence interval for the coefficient of education. Because validity of the exclusion restrictions is permitted, this interval overlaps with the TSLS interval. It yields the extremely wide and thus uninformative interval [-0.25, 0.24] for the direct effect of an additional year of schooling. Excluding *private* from this analysis (not presented in the Figures) yields the narrower but still very wide interval [-0.19, 0.20].

Adopting the rather wide interval [0, 0.4] for ρ_1, KLS produces for the direct effect θ_1 of education a value in the interval [-0.04, 0.04], whereas (taking into account that years at school do not accumulate experience, although increasing age) for education plus age

the effect $\beta_1 = \theta_1 + \theta_2$ is estimated to be in [-0.01, 0.08]. For mildly positive ρ_1 values the bottom row of graphs in Figure 4.3.2 shows that heteroskedasticity does not seem a problem, whereas the RESET tests – although less reassuring – do not strongly reject the specification either. Hence, there is no convincing evidence that KLS inference on model (4.8) is untenable, whereas there surely is for TSLS inference on model (4.7).

[Figure 4.3.3 here]

Therefore, it seems quite likely from Figure 4.3.3 that the earlier TSLS findings on the direct plus indirect effects of the controls *black*, *south* and *urb* are all strongly biased towards zero. According to the KLS findings their actual effects are much more pronounced, at the expense of the direct effect of years of education.

5. Conclusions

In this study, we demonstrate that empirical instrumental variables based findings will often be surrounded by serious doubts. Whether or not instruments are really valid cannot be assessed positively by unambiguous instrument-based data analysis, whereas we showed that mildly invalid instruments devastate the quality of inference. Irrespective of the validity of the instruments used, instrument-based inference is poor anyhow when instruments are weak. Then standard confidence intervals are over-optimistic, and more sophisticated weak-instrument robust confidence intervals are generally extremely wide and therefore often meaningless for practical decision making.

We highlight here an additional fundamental problem. This occurs in models where instrumental variables are being used to overcome omitted variables problems. Such studies have to be defended on the basis of theoretical arguments supporting the validity of the proposed external instruments. These are required to have no direct effect on the dependent variable, but a substantial indirect effect via the regressors for which one seeks consistent estimators for their direct causal effects. Any further regressors are just used as controls, in order to mitigate the complexity of the omitted component of the model. If the candidate external instruments have no direct effect on the dependent variable

indeed, this does not yet guarantee that they are valid instruments in the underspecified model. The additional requirement for that, which is usually not being discussed in most applications, is that these external instruments and the omitted explanatory variables are mutually uncorrelated, or if they are correlated, that this is just due to both having an association with the included control variables. Any association they may both have with any other variables renders the external instruments invalid. Hence, proper exclusion of the instruments from the fully specified model is insufficient; consistent estimation of the direct effects of the endogenous regressors in the model with omitted regressors requires extra arguments for validity of the instruments in the underspecified model. If the external instruments are valid indeed, the resulting estimators of the coefficients of the control variables will represent the sum of their direct effect and their indirect effect through the omitted variables.

So, the whole issue whether instrumental variables based inference is worthwhile in this context boils down to checking the following four aspects: (A) for which explanatory variables does one desire inference exclusively on their direct effect on the dependent variable; (B) are sufficient candidate external instruments available for which one can argue that they have no direct effect on the dependent variable; (C) both these candidate external instruments and the omitted regressors should not depend on the same causal factors, apart from the control variables of the model; and (D) partialling out any effects from the controls, the effect of the candidate external instruments on the endogenous regressors should have a magnitude that will lead to sufficiently efficient inference.

Aspects (B) and (C) cannot directly be examined by statistical methods, because this would require observations on the omitted regressors. A Sargan test, which is only available when one has more candidate external instruments than endogenous regressors, is really not equipped to provide a decisive judgement regarding (B) and (C) as we demonstrate in this study by simulation.

Therefore, an approach avoiding the use of instrumental variables all together seems most welcome. However, also the instrument-free approach laid out and illustrated here is certainly not free from hurdles. It requires to adopt numerical bounds on the correlation of the endogenous regressors and the unobserved model error. In a badly

specified model, in principle all regressors may be correlated with the disturbance term. When one is ignorant about these specification failures it seems impossible to make useful assumptions on the likely numerical range of the actual endogeneity correlations. In the present study, we have demonstrated the instrument-free approach when allowing for just one single endogenous regressor. Young (2019), surveying many journal volumes, reports that in about 90% of the articles presenting instrumental variables based inferences, the models concerned do just have one endogenous regressor. Hence, many practitioners consider this to be the most relevant case, although it is evident that simultaneous equations models will often have many more endogenous regressors.

On the other hand, in models with omitted variables problems, our analysis shows that just allowing for one endogenous regressor is vindicated when one focusses on the estimation of the direct causal effect of just one of its explanatories at a time. For such cases we could demonstrate for two classic empirical data sets that by the KLS technique, over a very wide range of possible endogeneity correlation values, misspecification test statistics can be presented which examine possible failures of the model in particular dimensions, including the wrong exclusion of controls, previously unavailable. In a KLS context, the interpretation of a (non-)rejection of such tests is reasonably straightforward, because it cannot be blurred by the possible use of invalid instruments. Therefore, a KLS-based test for omitted regressors, possibly cast into the special form of missing interactions or improper functional form (RESET), or geared to detect heteroskedasticity or serial correlation[6], may produce more solid evidence on the adequacy of adopted model assumptions than can be generated by instrument-based techniques. In the latter context, no misspecification test procedure can unequivocally disentangle whether it are the instruments, the model specification, or both, which require respecification.

One of the arguments put forward by Hansen (2019, p.434) when advising practitioners to use a very small significance level when interpreting the Sargan test of overidentifying restrictions is that the occurrence of rejections should be limited, simply because it is not at all clear what one should do when the Sargan test rejects. In our

[6]KLS cannot just be applied to cross-section data, but also to econometric time-series models, see Kiviet (2020c).

opinion such clarity can be provided now: Apply KLS, and do so, too, when the Sargan test does not reject.

Supplementary material

The three appendices of this paper are available as one separate document. Code for the simulations (Matlab) and for the applications (Stata) can be obtained from the authors.

References

Andrews, D.W.K., Stock, J.H., 2007. Inference with weak instruments. In: Blundell, R., Newey, W.K., Persson, T. (Eds.), *Advances in Economics and Econometrics, Theory and Applications*: Ninth World Congress of the Econometric Society. Cambridge University Press, Cambridge.

Andrews, I., Stock, J., Sun, L., 2019. Weak instruments in IV regression: Theory and practice. *Annual Review of Economics* 11, 727–753.

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 91, 444-455.

Angrist, J.D., Krueger, A.B., 2001. Instrumental variables and the search for Identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* 15, 69–85.

Bontemps, C., Magnac, T., (2017). Set identification, moment restrictions, and inference. *Annual Review of Economics* 9, 103-129.

Bound, J., Jaeger, D., 2000. Do compulsory attendance laws alone explain the association between earnings and quarter of birth? *Research in Labor Economics* 19, 83-108.

Bound, J., Jaeger, D., Baker, R., 1995. Problems with Instrumental Variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *Journal of the American Statistical Association* 90, 443-450.

Card, D., 1995. Using geographic variation in college proximity to estimate the return to schooling. In: *Aspects of Labor Market Behavior: Essays in Honour of John Vanderkamp.* Editors: Christofides, L.N., Grant, E.K., Swidinsky, R., Toronto: University of Toronto Press.

Card, D., 1999. The Causal Effect of Education on Earnings, in: *Handbook of Labor Economics*, Volume 3A, ed. by Orley Ashenfelter and David Card. Amsterdam and New York: North Holland.

Card,.D., 2001. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69, 1127-1160.

Clarke, D., Matta, B., 2018. Practical considerations for questionable IVs. *The Stata Journal* 18, 663-691.

Conley, T.G., Hansen, C.B., Rossi, P.E., 2012. Plausibly exogenous. *The Review of Economics and Statistics* 94, 260-272.

Hansen, B.E., 2019. *Econometrics.* Can be downloaded at https://www.ssc.wisc.edu/~bhansen/econometrics/

Harding, M., Hausman, J., Palmer, C. (2016). Finite sample bias corrected IV estimation for weak and many instruments", Essays in Honor of Aman Ullah (*Advances in Econometrics*, Vol. 36), Emerald Group Publishing Limited, 245-273.

Kiviet, J.F., 2017. Discriminating between (in)valid external instruments and (in)valid exclusion restrictions. *Journal of Econometric Methods* 6, 1-9.

Kiviet, J.F., 2020a. Testing the impossible: identifying exclusion restrictions. To appear in *The Journal of Econometrics,* https://doi.org/10.1016/j.jeconom.2020.04.018.

Kiviet, J.F., 2020b. Instrument-free inference under confined regressor endogeneity; derivations and applications. *Mimeo.*

Kiviet, J.F., 2020c. Causes of haze and its health effects in Singapore; a replication study. *Mimeo.*

Kiviet, J.F., Niemczyk, J., 2012. The asymptotic and finite sample (un)conditional distributions of OLS and simple IV in simultaneous equations. *Journal of Computational Statistics and Data Analysis* 56, 3567-3586.

Kiviet, J.F., Phillips, G.D.A., 1993. Alternative bias approximations in regressions

with a lagged dependent variable *Econometric Theory* 9, 62-80.

Kraay, A., 2012. Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach. *Journal of Applied Econometrics* 27, 108-128.

Kripfganz, S., Kiviet, J.F., 2020. kinkyreg: Instrument-free inference for linear regression models with endogenous regressors. Community-contributed Stata program. *Mimeo* in preparation.

MacKinnon, J.G., Smith, A.A., 1998. Approximate bias correction in econometrics. *Journal of Econometrics* 85, 205-230.

Murray, M.P., 2006. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives* 20, 111-132.

Nelson, C.R., Startz, R., 1990. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58, 967-976.

Nevo, A., Rosen, A.M., 2012. Identification with imperfect instruments. *The Review of Economics and Statistics* 94, 659-671.

Parente, P.M.D.C, Santos Silva, J.M.C., 2012. A cautionary note on tests of overidentifying restrictions. *Economics Letters* 115, 314-317.

Rosenzweig, M.R., Wolpin, K.I., 2000. Natural experiments in economics. *Journal of Economic Literature* 37, 827–874.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 557-586.

Stock, J.H., Wright, J.H., Yogo,. M., 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20, 518-529.

Young, A., 2019. Consistency without inference: Instrumental Variables in practical application. Unpublished manuscript.

Figure 4.2.1 KLS results on model specification (12.96) for the Angrist-Krueger (1991) data

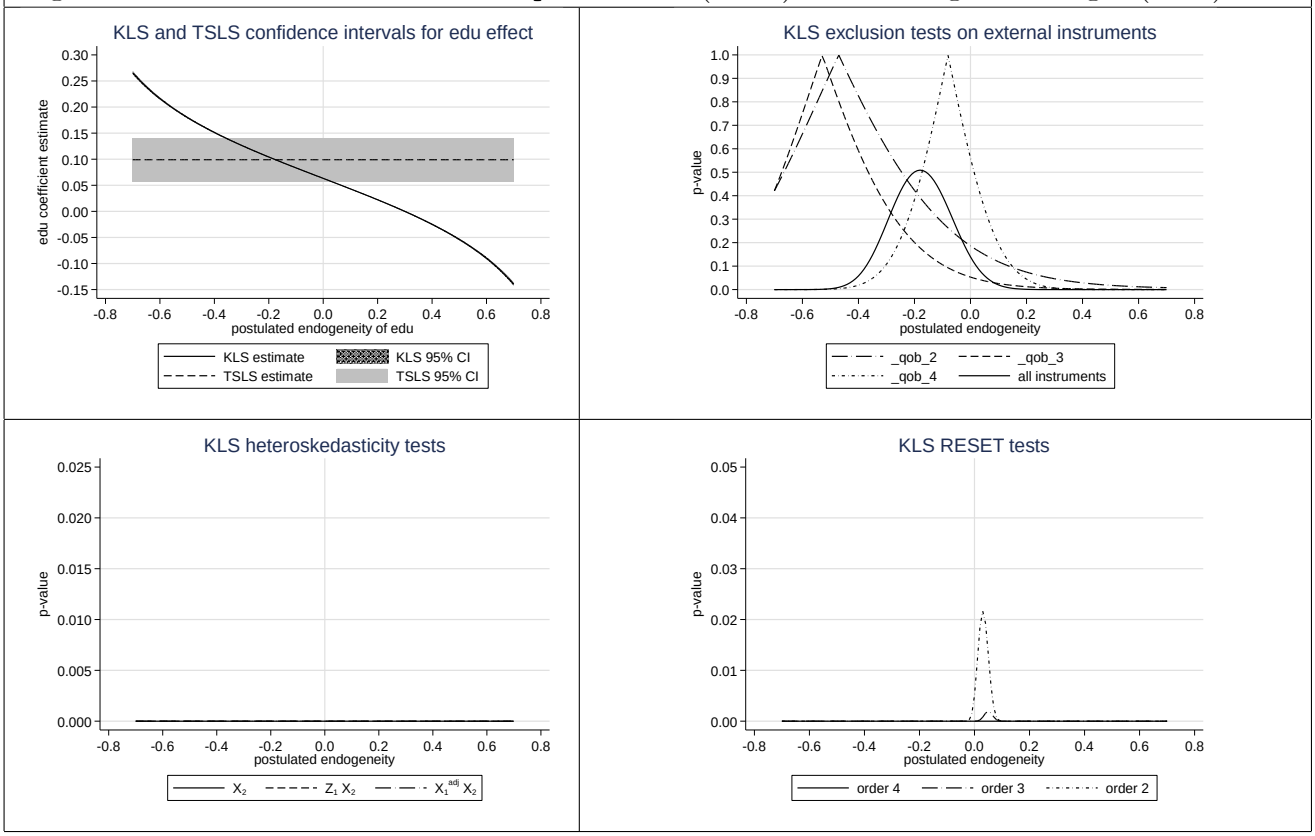

Figure 4.2.2 KLS results on model specification (4.3) for the Angrist-Krueger (1991) data

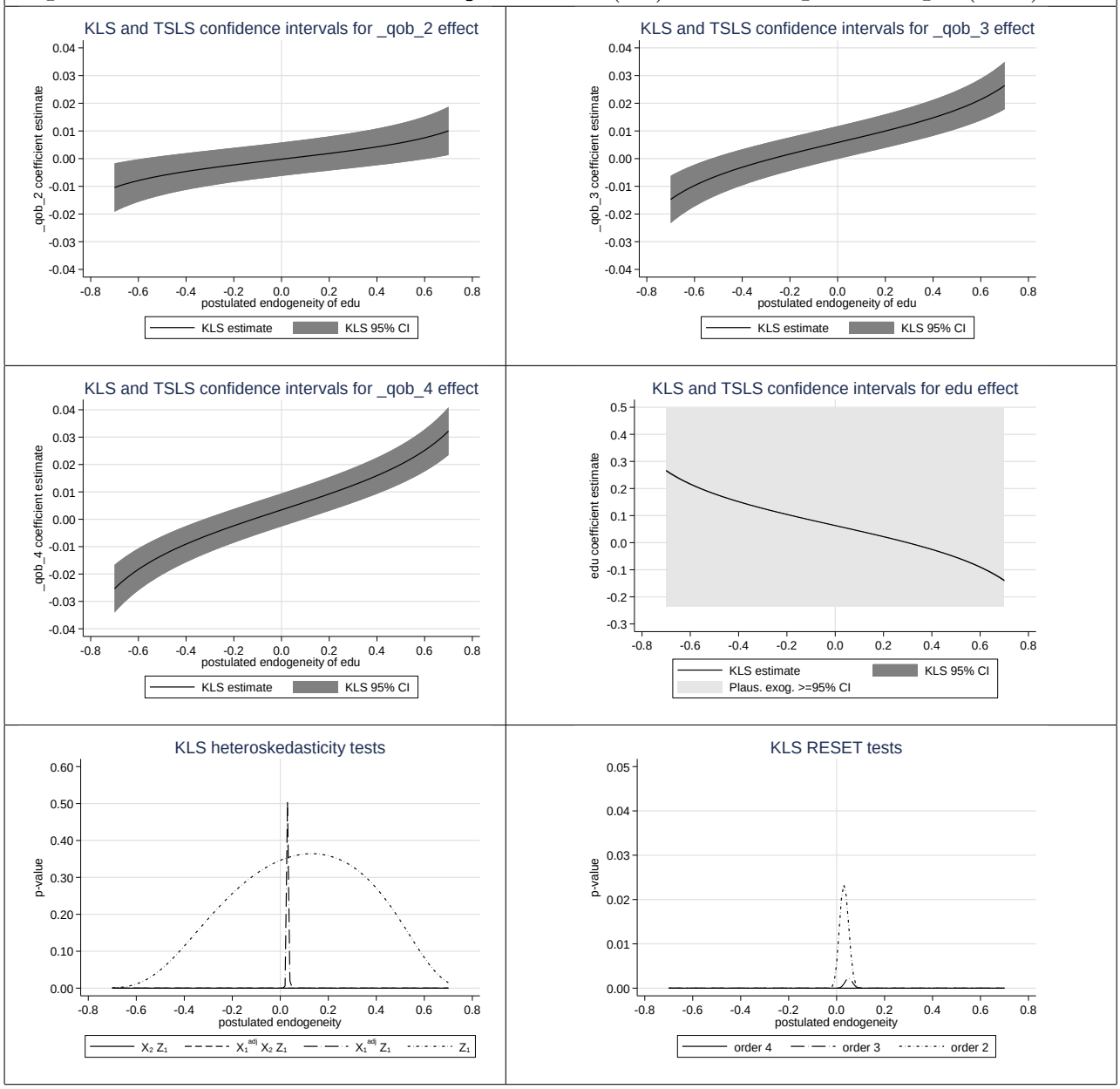

Figure 4.3.1 KLS results on model specification (4.7) for the Card (1995) data

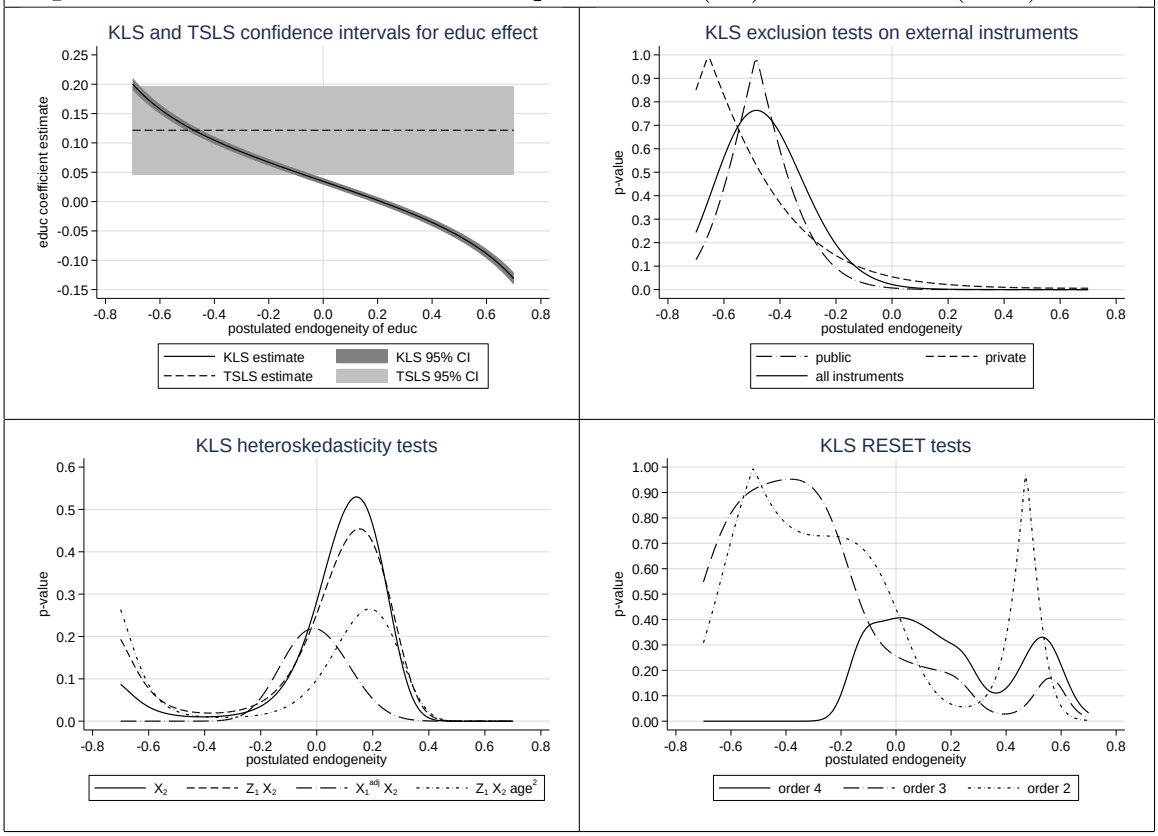

Figure 4.3.2 KLS results on model specification (4.8) for the Card (1995) data

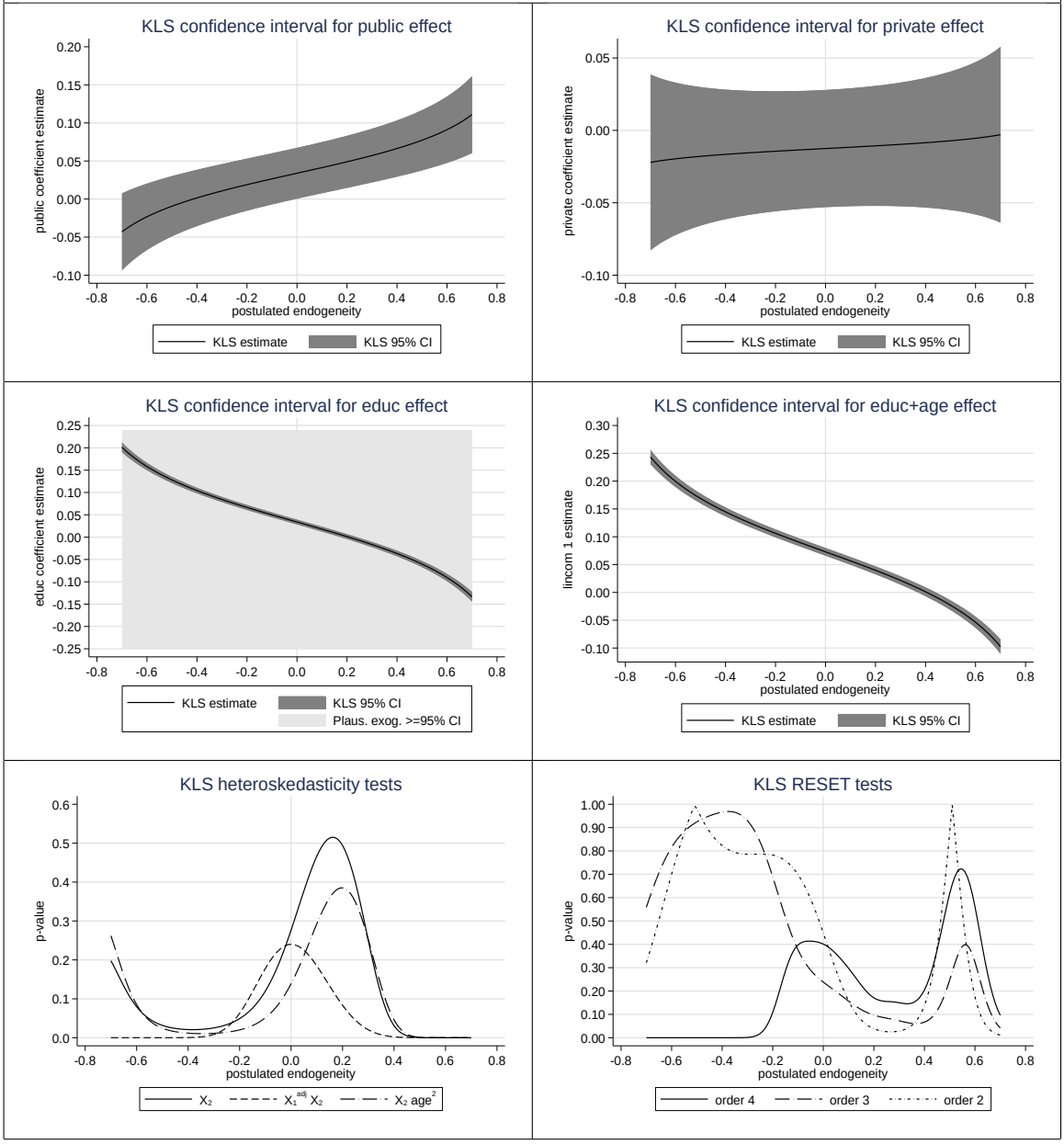

Figure 4.3.3 KLS results on model specification (4.8) for the Card (1995) data