

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

Jan F. Kiviet^{*}

Sebastian Kripfganz[†]

Abstract

For two classic case studies in labor economics, it is shown how recently developed instrument-free methods yield remarkably narrow asymptotically valid confidence intervals for regression coefficients. These methods achieve set-identification through adopting credible ranges for the correlation between endogenous regressors and model errors. They also provide more decisive evidence on (in)validity of exclusion restrictions than Sargan-Hansen over-identification tests. The latter has been shown to often not detect serious instrument invalidity, whereas instrumental variables based inference (also identification robust variants) suffers severely even from mild invalidity, especially for weak instruments.

JEL Classification: C10, C26, I26, J31

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

1 Introduction

With the purpose to investigate whether recently developed new instrument-free inference techniques reinforce the credibility of the original results or provide new insights, we apply them to data earlier used in two classic case studies in labor economics (Angrist and Krueger, 1991; Card, 1995). To date, it is still current best practice to use instrumental

^{*}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.

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

variables based techniques for such analyses. We conclude from the results 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 adopting instruments. Our findings for these case studies raise doubts about the validity of the exclusion restrictions and about the reliability of over-identification tests for the detection of such a violation.

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 variables 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 microeconomic 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 stan-

dard form of 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, if at all, 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 often 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) mostly held on to employing instrument-based techniques. Kraay (2012) used Bayesian methods allowing for a certain degree of instrument 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

¹For (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).

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). It pursues identification by making point or interval assumptions on the actual numerical value of the correlations between the regressors and the error term. Set identification is achieved, as defined in Bontemps and Magnac (2017), provided the specified intervals cover the true correlations. 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.

The KLS approach is related to alternative instrument-free identification strategies that are used for sensitivity analysis. In the absence of reliable exclusion restrictions, Altonji et al. (2005) obtain bounds on the estimates by making assumptions on the relative importance of selection on observables versus selection on unobservables. Effectively, their approach is bounding the correlation between the error terms in the outcome and the selection equation. Oster (2019) obtains a consistent set-identified estimator for bias-adjusted treatment effects by imposing similar bounds on the relative degree of selection on observables and unobservables, and by bounding the coefficient of determination that could be hypothetically achieved if all unobserved control variables were included in the regression. The underlying logic is similar to that of KLS, namely to achieve interval identification by imposing bounds on some aspects of the relationship between observed and unobserved model characteristics. Choosing these bounds requires additional prior

knowledge or assumptions. Both Altonji et al. (2005) and Oster (2019), who just consider models with one endogenous regressor, find it reasonable to assume that the relationship between the endogenous regressor and the unobserved variables is not stronger than its relationship with the observed control variables. However, in practice the resulting bounds can easily become too wide for informative inference. KLS bounds leading to sufficiently narrow confidence intervals might be easier to justify, or at least help to examine the plausibility of the results obtained with alternative approaches, also because KLS unlocks an extra layer of endogeneity-robust misspecification tests. Although destructive for the specifications originating from the classic studies, KLS can and should be turned into a constructive tool for discovering better models from extended data sets.

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. Their assessment, against which our empirical findings will be judged, are obtained from a small scale simulation study. All its details are presented in Appendices A and B, separately available as Supplementary material. Section 4 first analyzes crucial issues on the endogeneity and instrumentation problems affecting the empirical analysis of models suffering from omitted regressors, and of earnings equations in particular. Most studies leave these issues implicit; we provide full technical details on them in Appendix C. Next, we contrast empirical results from instrument-based and instrument-free approaches, namely standard TSLS and KLS, and the approaches put forward by Conley et al. (2012) and by Oster (2019). We do this for the two classic studies on the causal effect of education on earnings by Angrist and Krueger (1991) and Card (1995). Finally, Section 5 concludes.

2 The essentials of 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 point or set identification 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.

For a biased but consistent estimator, the order of magnitude of a consistent estimator's 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, proved 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.

For clarification, we provide here some 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 $i = 1, \dots, n$ we have

$$u_i \sim (0, \sigma_u^2), \quad x_i \sim (0, \sigma_x^2), \quad \text{with } E(x_i u_i) = \rho_{xu} \sigma_x \sigma_u. \quad (1)$$

Applying ordinary least-squares (OLS) yields the estimators

$$\hat{\beta}_{OLS} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2, \quad \hat{u}_i = y_i - \hat{\beta}_{OLS} x_i, \quad \text{and } \hat{\sigma}_u^2 = n^{-1} \sum_{i=1}^n \hat{u}_i^2, \quad (2)$$

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

by the estimators

$$\hat{\beta}_{KLS}(\rho_{xu}) = \hat{\beta}_{OLS} - \rho_{xu} \sqrt{\frac{\hat{\sigma}_u^2(\rho_{xu})}{n^{-1}\sum_{i=1}^n x_i^2}}, \quad (3)$$

$$\hat{\sigma}_u^2(\rho_{xu}) = \hat{\sigma}_u^2/(1 - \rho_{xu}^2), \quad (4)$$

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

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

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

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 $\sigma_u^2/\sum_{i=1}^n x_i^2$, which is invariant regarding ρ_{xu} , and is consistently estimated by $\hat{\sigma}_u^2(\rho_{xu})/\sum_{i=1}^n x_i^2$.

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

$$[\hat{\beta}_{KLS}(\rho_{xu}) - \beta_0]/[\widehat{Var}(\hat{\beta}_{KLS}(\rho_{xu}))]^{1/2} \quad (6)$$

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

$$\hat{\beta}_{KLS}(\rho_{xu}) \pm \zeta_{1-\alpha/2} [\widehat{Var}(\hat{\beta}_{KLS}(\rho_{xu}))]^{1/2}, \quad (7)$$

where ζ_p is the p^{th} quantile of the standard normal distribution. Such unfeasible (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.

At first sight KLS seems unfeasible, because ρ_{xu} is generally unknown in practice. However, KLS inference can 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. Similarly, in the framework of Altonji et al. (2005) and Oster (2019), the sensitivity of the results can be assessed with respect to a varying degree of proportionality regarding the selection on unobservables relative to observables, while any particular choice for the proportionality parameter may be hard to motivate. Importantly, any one of the approaches might be informative about the plausibility of the assumptions required for one of the other approaches, thus making them complements rather than substitutes.

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

We ran a small scale Monte Carlo study to develop some useful intuition regarding the (lack of) qualities of TSLS and KLS under relevant circumstances. It is based on simulated though rather typical cross-section data for a single relationship with just one

endogenous regressor and an intercept, for which two candidate external instruments are available. Samples of size 50, 250 and 2500 have been analyzed. We focussed on non-extreme cases, avoiding seriously weak or exceptionally invalid instruments. We compared the median bias and interquartile range of TSLS and KLS estimators, examined the sensitivity of TSLS regarding using invalid instruments, assessed the effectiveness of over-identification testing for deciding on the (in)validity of instruments, and investigated the effects on KLS of using (in)valid assumptions on the actual degree of endogeneity. All details on these experiments can be found in Appendices A and B; see also Kiviet and Kripfganz (2021). In this section we just report the major findings.

For standard linear models, the over-identification restrictions test is the Sargan (1958) test. For its correspondences with testing exclusion restrictions, and its implicit adoption of just-identifying restrictions, see Parente and Santos Silva (2012) and Kiviet (2017). In the literature this test is often blamed for over-rejection in finite samples. Therefore, some authors advise practitioners to use the test at a very low nominal significance level; see Hansen (2022, Ch.12). On the other hand one could argue in favor of testing at a very high nominal significance level, because an insignificant value of the test is used in practice to accept the null hypothesis of validity of all instruments. So, the primary worry being to fail to reject invalid instruments (commit type II errors) and not so much to limit type I errors (wrongly rejecting valid instruments) one might decide 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%, or just 1% as Hansen (2022) suggests!

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 (more or less mandatory) Sargan test (and its generalization the J -test) and the consequences for inference on model coefficients of type II errors of these tests. For our models based on i.i.d. observations 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 $0.05 \leq \alpha \leq 0.5$ (so 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 other 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 seriously lacks power.

Of course we find that 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 is TSLS 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. When instruments are invalid, just and over-identified IV/TSLS are 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 TSLS, 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 TSLS is serious and the interquartile range much larger than for KLS. The findings on the relative width of their actual interquartile ranges demonstrate that also so-called identification-robust IV/TSLS inference must often be less accurate than feasible 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 biased as well. Therefore, we also simulated for KLS its median (which proved to be invariant with respect to the sample size) and interquartile range when using an erroneous ρ_{xu} . The results made us conclude that the vulnerability of KLS to moderate errors in

bounding ρ_{xu} , 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). We also explore whether the sensitivity analysis suggested by Oster (2019) can add further value. Moreover, we will employ KLS-based instrument-free misspecification tests. These are found to detect model failures which previously remained unnoticed.

Although we are aware that both the Angrist-Krueger and the Card analyses have triggered a rich literature with many fruitful suggestions for making different use of the classic data sets or of exploiting further relevant variables, we nevertheless limit ourselves in this replication study to just a few of the empirical results from the well-known original published articles. Our selection here is in fact strongly influenced by the particular illustrations from these articles presented in the very recent textbook by Hansen (2022). These classic IV/TSLS results suffice in our opinion to illustrate the various pros and cons of the more recently developed instrument-based and instrument-free approaches.

Before we address the two classic studies, in the next subsection we first give 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. Appendix C provides a more formal treatment of these issues.

4.1 Endogeneity and instrumentation when regressors have been omitted

The applications to be considered are characterized by the following. The causal 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

$$y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon, \quad (8)$$

where the X_j are $n \times K_j$ matrices and y and ε are $n \times 1$ vectors with $E(\varepsilon | X_1, X_2, X_3) = 0$, so all regressors are exogenous with respect to ε . 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 $\text{rank}(X) = K_1 + K_2 > 0$. 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

$$\gamma = X_3\beta_3 \neq 0, \text{ where } E(\gamma | X) = X_1\phi_1 + X_2\phi_2. \quad (9)$$

Hence, we allow that X_3 is associated with X_1 and X_2 . Substitution of (9) into (8) and defining $\eta = \gamma - E(\gamma | X)$, with $E(\eta | X) = 0$, gives

$$y = X_1(\beta_1 + \phi_1) + X_2(\beta_2 + \phi_2) + (\eta + \varepsilon), \text{ with } E(\eta + \varepsilon | X) = 0.$$

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 a consistent estimator 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 (8) with the applications to follow. In these, the dependent variable y contains observations on the log 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: intelligence, having special skills, and notions expressing appearance, upbringing, and charm. In model (8), 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 , where $E(\gamma | X_2) = X_2\phi_2^*$. 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 (8). 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 allocating more regressors from X_2 to X_1 (so aiming to estimate the direct causal effect of more explanatory variables 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 (2022, Ch.12) 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, he treats the regressor education in years (*edu*) as 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. That now the coefficients of the latter will no longer represent direct effects should be (but usually is not) mentioned in this literature. The reduced form equation for *edu* yields an *F*-test of 31 on the joint significance of the three quarter of birth (*qob*) dummies, so they seem

sufficiently strong indeed based on common practice.

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. Any further relevant omitted regressors that are correlated with an external instrument 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 used by Hansen (2022) in his textbook, and also examined in Conley et al. (2012, p.269), we present KLS results² 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),

²The results for empirical data have all been obtained by Stata, employing for KLS the *kinkyreg* Stata program contributed to the Stata community and documented and illustrated in Kripfganz and Kiviet (2021).

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.³ 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 $\rho_1 \geq 0$. These KLS inferences are in sharp contrast with the established TSLS conclusions.

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. Buckles and Hungerman (2013) provide evidence that this is indeed the case for family background

³Further substantiation of the adopted sign and numerical bounds of ρ_1 can be based on formula (A.19) of Kiviet (2020b).

variables.

The bottom two panels of Figure 4.2.1 demonstrate that the considered model fails the various KLS-based misspecification tests regarding heteroskedasticity⁴ and RESET⁵ for any value of ρ_1 . Hence, we have to conclude that we need to be cautious about making inference on β_1 from the top-left panel.

[Figure 4.2.2 here]

The rejection of the exclusion restrictions could also inspire to adhere to 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 ψ in the augmented model

$$y = X_1\beta_1 + X_2\beta_2 + Z_1\psi + \varepsilon, \quad (10)$$

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 . We can obtain empirical support for assumptions on ψ 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 (10). 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 (10) and also by using the plausibly exogenous approach. Note that the latter interval is invariant regarding ρ_1 . It overlaps the TSLS interval presented in Figure 4.2.1 (because the chosen intervals do not exclude $\psi = 0$), it has asymptotic confidence coefficient exceeding 95%, and is so wide that it is of little practical use.⁶ 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

⁴Here 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 .

⁵Here the joint significance is tested of the additional regressors, when the model is augmented by $\hat{y}_i^2, \dots, \hat{y}_i^d$, where the integer order is $d \geq 2$ and \hat{y}_i is the estimated exogenous component of y_i .

⁶Using the same data, Conley et al. (2012) produce similar results, where all three *qob* coefficients are supposed to have a value in $[-\psi, \psi]$ for $\psi \in [0, 0.02]$. Focussing on $\psi = 0.01$ they conclude that, using their method, the data are essentially uninformative about the returns to schooling.

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 (10) very strongly. On the other hand, given 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.

To obtain further insights on the sensitivity of the results, we might also employ the approach put forward by Oster (2019). In line with her suggestions, we start by conservatively assuming that there is equal selection on the observables and the unobservables (also suggested by Altonji et al., 2005), and that hypothetically including all unobserved controls would fully explain the variation in the outcome variable (thus not allowing for measurement error in wages). Formally, this implies in our notation

$$\delta = \frac{E(\gamma_i^* X_{i1})}{E(\gamma_i^*)^2} \frac{E(X_{i2}\beta_2^*)^2}{E(X_{i2}\beta_2^* X_{i1})} = 1$$

and $Var(\varepsilon_i) = 0$, based on the model in (4.1) and (4.2), and where $\beta_2^* = \beta_2 + \phi_2^*$. ϕ_2^* are the coefficients in a linear prediction of the unobserved component γ on the observed control variables X_2 , and γ^* is the corresponding prediction error that is orthogonal to X_2 ; see Appendix C. A practical complication of the Oster (2019) estimator is that it can yield multiple solutions. Only one of them is consistent, but it may not always be obvious which of them it is. Computed with Oster's *psacalc* Stata command, we obtain -0.137 and 0.252 as the candidate solutions for the return to schooling under the above assumptions. Compared to the KLS results, neither of them appears plausible as they correspond to the KLS estimates for an either extremely positive or extremely negative

ρ_1 . Oster (2019) proposes to select the solution under which the omitted-variables bias does not change the direction of the covariance between the endogenous regressor and the observed control variables. Here, this would be -0.137, which tends towards zero when the maximum R-squared, R_{max} , from the hypothetical regression with all control variables is lowered. The second solution would become even larger and thus more implausible. KLS can thus help to pick the correct solution for the Oster (2019) estimator. For $R_{max} = 0.8$ or $R_{max} = 0.5$, the bias-adjusted estimator of the returns to schooling becomes -0.08 and -0.004, respectively. Those estimates still correspond to a relatively large ρ_1 , and the negative effect sign is hardly convincing. The sign eventually flips when we also lower the bound on the proportionality factor, δ , for the selection on observables to the selection of unobservables. With $R_{max} = 0.5$, an effect size of 0.035 can be reached by lowering δ from 1 to 0.5, i.e. making selection on observables twice as important as selection on unobservables. Yet, the choices for these tuning parameters are quite arbitrary. In this example, the KLS estimates help to narrow the bounds for the Oster (2019) approach, but the latter hardly helps to improve inference without informative prior knowledge on δ and R_{max} .

4.3 Card (1995)

Card too examines the individual wage equation. 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 (with non-robustified standard errors in parentheses) 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 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 than one endogenous regressor –although possible– is a bit cumbersome too, we will choose a different road. Because

$$exper = age - educ - 6, \quad (11)$$

the model

$$lwage = \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \dots + u \quad (12)$$

implies

$$\begin{aligned} lwage &= \beta_1 educ + \beta_2(age - educ - 6) + \beta_3(age - educ - 6)^2 + \dots + u \\ &= (\beta_1 - \beta_2 + 12\beta_3)educ + (\beta_2 + 12\beta_3)age + \beta_3 age^2 - 2\beta_3 age \times educ + \dots + u \\ &= \theta_1 educ + \theta_2 age + \theta_3 age^2 + \theta_4 age \times educ + \dots + u. \end{aligned} \quad (13)$$

If $\beta_3 = 0$, then $\theta_3 = \theta_4 = 0$ and $\beta_1 = \theta_1 + \theta_2$. Hence, inference on β_1 of (12) 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 (13). 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 (13) yields (omitting presentation of the intercept):

$$lwage = 0.094educ + 0.082age - 0.073age^2/100 - 0.101black - 0.099south + 0.108urb$$

(0.050)	(0.071)	(0.124)	(0.075)	(0.030)	(0.050)
[0.049]	[0.070]	[0.123]	[0.073]	[0.030]	[0.050]

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:

$$lwage = 0.121educ + 0.054age - 0.023age^2/100 - 0.060black - 0.085south + 0.082urb$$

(0.038)	(0.065)	(0.114)	(0.059)	(0.026)	(0.040)
[0.038]	[0.065]	[0.113]	[0.058]	[0.027]	[0.040]

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:

$$\begin{aligned}
 lwage = & 0.121educ + 0.041age - 0.060black - 0.085south + 0.082urb \quad (14) \\
 & (0.038) \quad (0.003) \quad (0.060) \quad (0.026) \quad (0.041) \\
 & [0.038] \quad [0.003] \quad [0.058] \quad [0.027] \quad [0.040]
 \end{aligned}$$

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 (14).

[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 (14) 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 the IV analysis, when just using the *public* college dummy as an instrument, residuals are obtained which have correlation -0.48 with the endogenous regressor *education*, 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 (14) 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 (14). 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:

$$\begin{aligned}
lwage = & 0.018 \textit{educ} + 0.039 \textit{age} - 0.212 \textit{black} - 0.129 \textit{south} + 0.171 \textit{urb} \\
& (0.003) \quad (0.002) \quad (0.018) \quad (0.016) \quad (0.017) \\
& + 0.041 \textit{public} - 0.012 \textit{private} \\
& (0.017) \quad (0.021)
\end{aligned} \tag{15}$$

We note that the former instruments are not both individually significant when $\rho_1 = 0.1$. Comparing with (14), KLS yields much smaller standard errors and substantially different coefficient values. More detailed results on model specification (15) 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 $0 \leq \rho_1 \leq 0.4$, 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. (2012) 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 (15) is untenable, whereas there surely is for TSLS inference on model (14).

[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.

Further sensitivity analysis for the sum of the two coefficients θ_1 and θ_2 in model (4.6) along the lines proposed by Oster (2019) is not feasible for this application, as it is only applicable to a single treatment effect. Reverting to specification (4.5) is not helpful either due to the link between *exper* and *educ*. Assuming that *age* and other control variables are not informative for the unobserved skills/ability component, we cannot disentangle the influence of ability on *educ* from that on *exper* (and other control variables). Put differently, after netting out *exper* from the unobservables, *educ* will be unrelated to the unobservables as well.

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 (2022), 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 correlation⁷, 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.

Especially in models with omitted variables it seems not unlikely that the errors may be affected by conditional heteroskedasticity, whereas the present KLS procedure pre-

⁷KLS cannot just be applied to cross-section data, but also to econometric time series models, see Kiviet (2020c).

supposes conditional homoskedasticity. The derivation of heteroskedasticity-consistent standard errors for KLS, although challenging, has not yet been undertaken. However, KLS heteroskedasticity testing can eventually be used to estimate a nonconstant skedasticity function, which may be used to weight the data. The results in Romano and Wolf (2017) suggest that aiming to find proper weights for the sample observations, in order to regain homoskedasticity, may contribute more to improving inference than just employing robustification naively to unweighted data. Hence, even if a KLS robustification formula were available, one should not use this as an excuse to not constructively approach any heteroskedasticity.

One of the arguments put forward by Hansen (2022) when advising practitioners to use a very small significance level when interpreting the Sargan test of over-identifying 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 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 applications (Stata) and the simulations (Matlab) can be obtained from the authors.

References

- Altonji, J.G., Elder, T.E., Taber, C.R., 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113, 151–184.
- 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 school-

ing 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.

Buckles, K.S., Hungerman, D.M., 2013. Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics* 95, 711-724.

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. Editors: Ashenfelter, O., Card, D. 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., 2022. *Econometrics*. Princeton University Press.

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. *Journal of Econometrics* 218, 294-316.
- Kiviet, J.F., 2020b. Causes of haze and its health effects in Singapore: a replication study. *The Singapore Economic Review*, 65, 1367-1387.
- Kiviet, J.F., 2023. Instrument-free inference under confined regressor endogeneity and mild regularity. *Econometrics and Statistics* 25, 1-22.
- Kiviet, J.F., Kripfganz, S., 2021. Instrument approval by the Sargan test and its consequences for coefficient estimation. *Economics Letters* 205, 109935.
- 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., 2021. kinkyreg: Instrument-free inference for linear regression models with endogenous regressors. *Stata Journal* 21, 772-813.
- 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.
- Oster, E., 2019. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics* 37, 187-204.
- Parente, P.M.D.C., Santos Silva, J.M.C., 2012. A cautionary note on tests of overidentifying restrictions. *Economics Letters* 115, 314-317.
- Romano, J.P., Wolf, M., 2017. Resurrecting weighted least squares. *Journal of Econometrics* 197, 1-19.
- Rosenzweig, M.R., Wolpin, K.I., 2000. Natural experiments in economics. *Journal of*

Economic Literature 37, 827–874.

Sargan, J.D., 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393-415.

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., 2022. Consistency without inference: Instrumental Variables in practical application. *European Economic Review* 147, 104112.

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

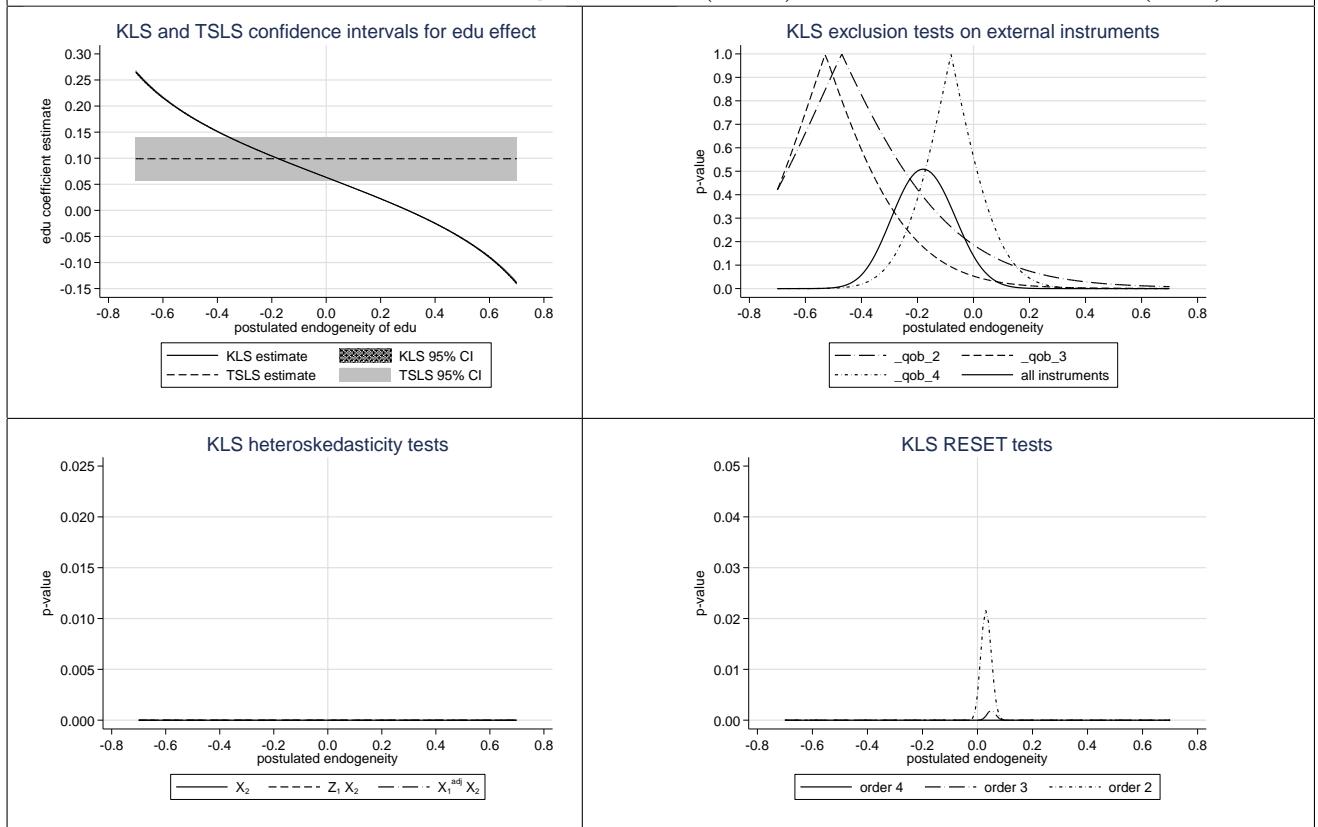


Figure 4.2.2 KLS results on model specification (4.4) for the Angrist-Krueger (1991) data

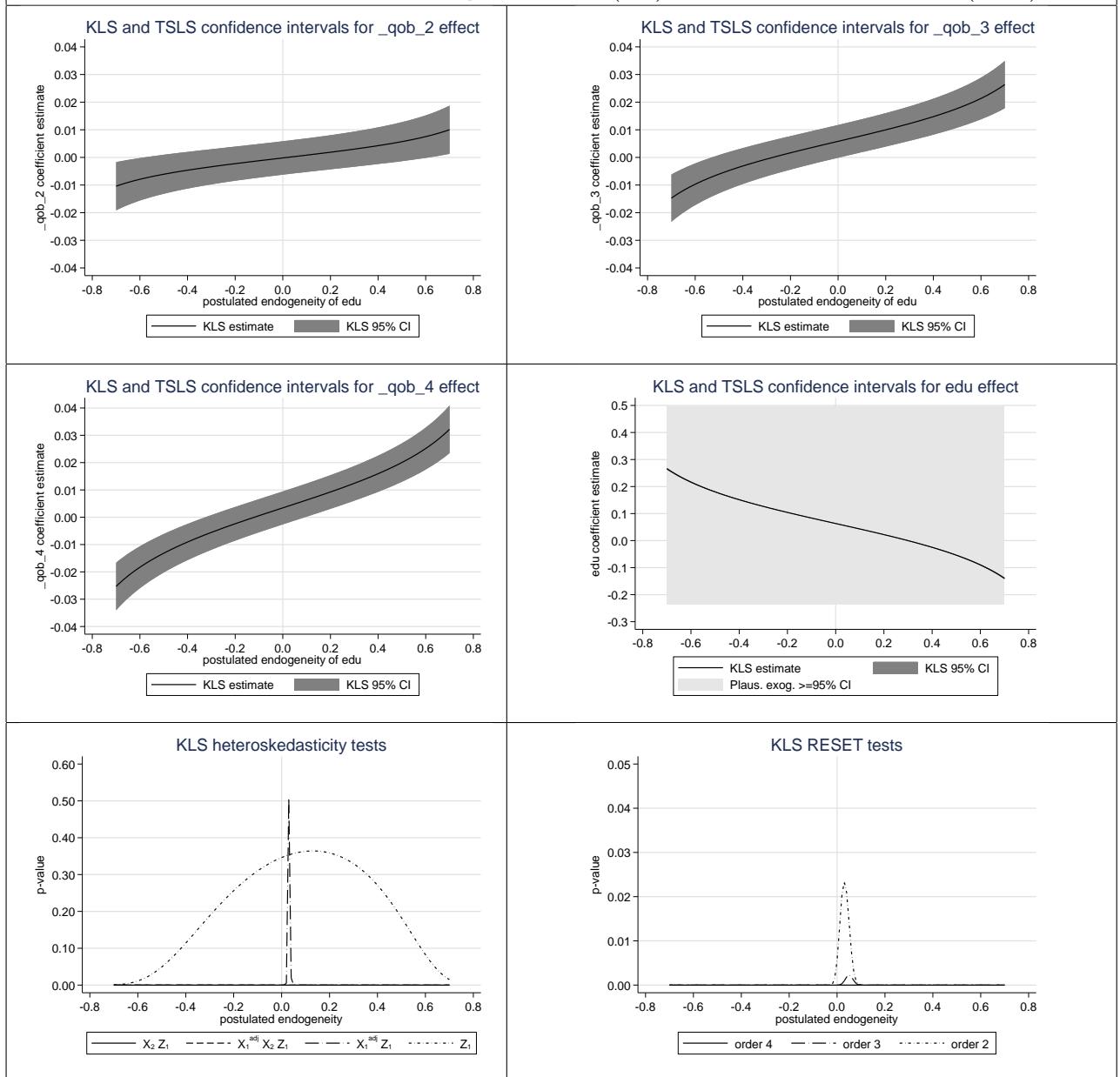


Figure 4.3.1 KLS results on model specification (4.9) for the Card (1995) data

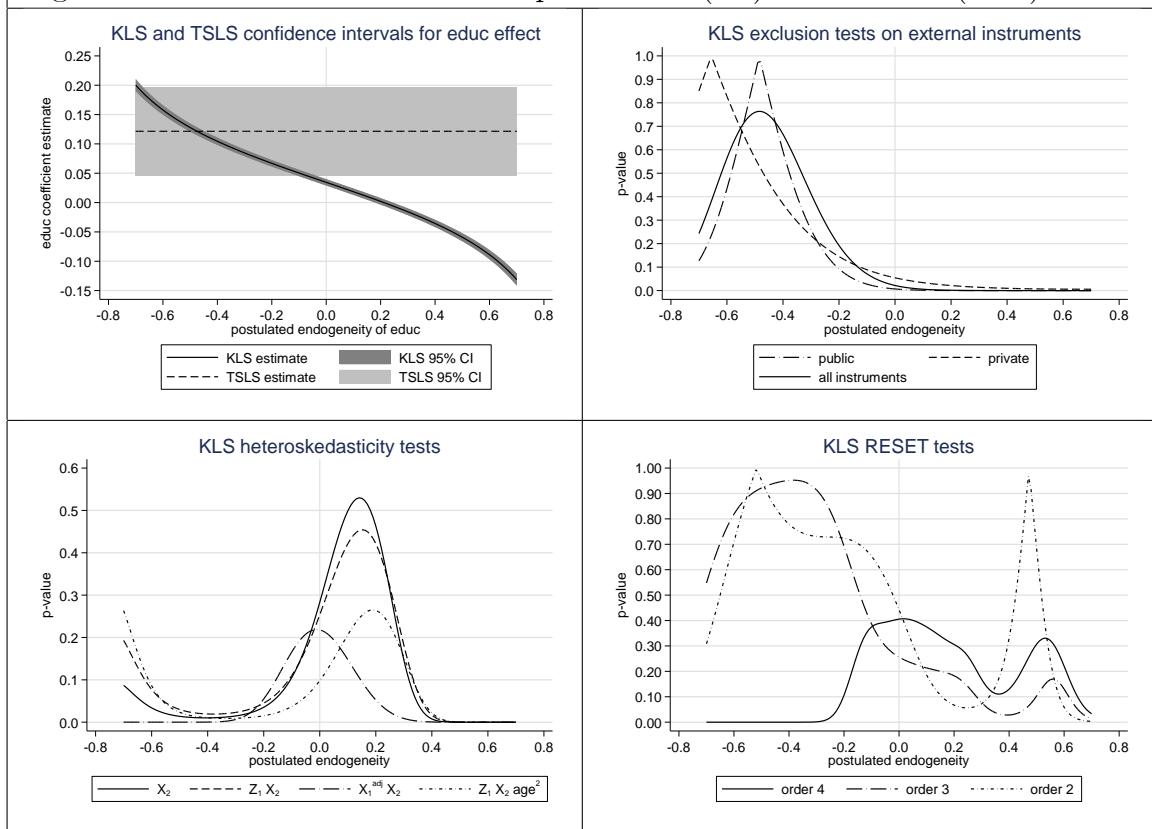


Figure 4.3.2 KLS results on model specification (4.10) for the Card (1995) data

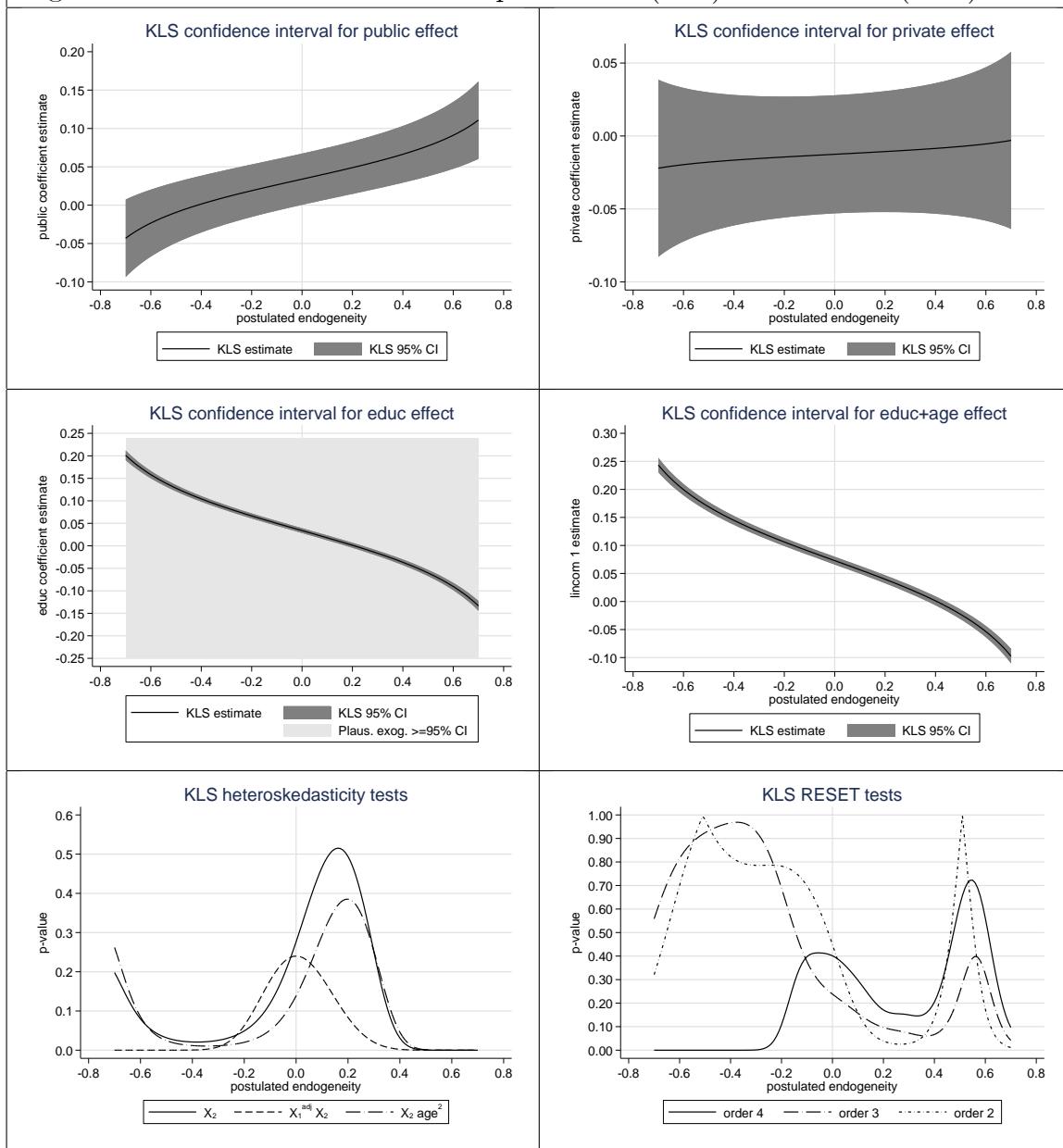
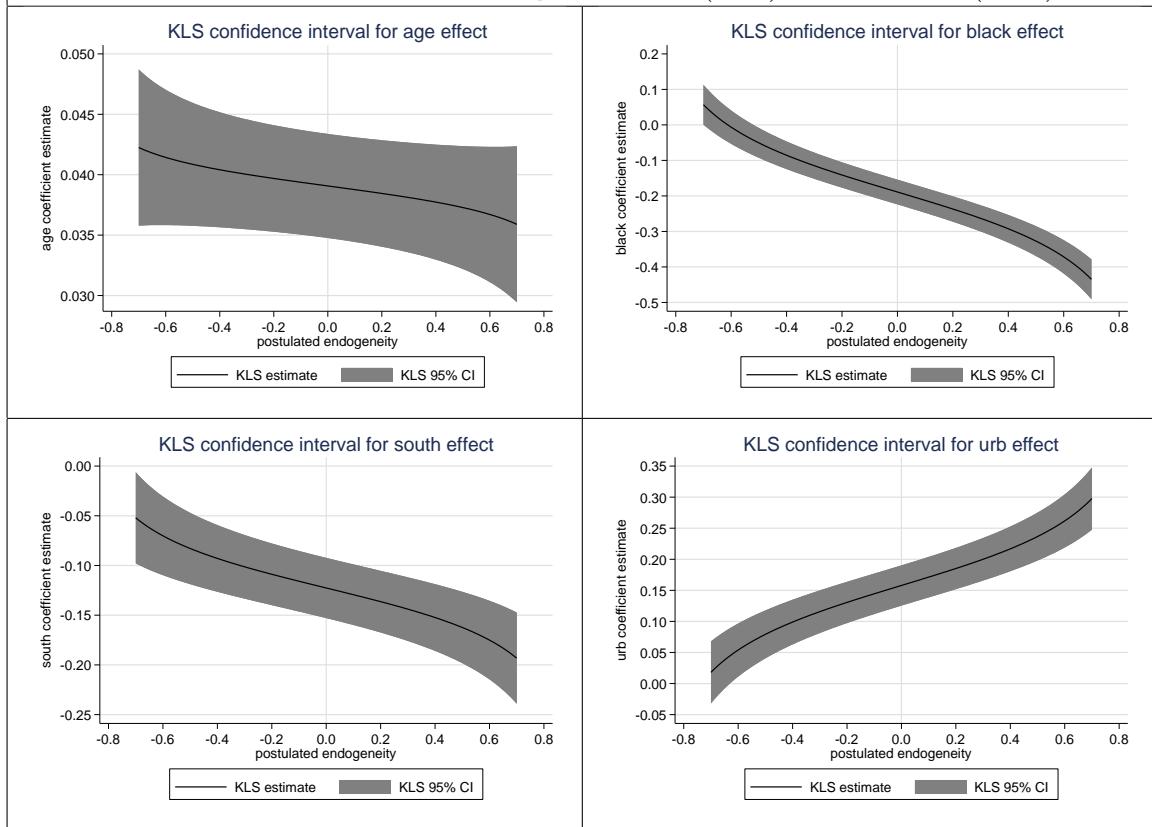


Figure 4.3.3 KLS results on model specification (4.10) for the Card (1995) data



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

Jan F. Kiviet

Sebastian Kripfganz

Appendices

The following appendices provide further background to the Monte Carlo simulation study and to the options to cope with omitted relevant regressors by TSLS and KLS:

Appendix A gives a detailed description of the graphically presented findings from the simulation study.

Appendix B provides full technical details on the chosen design of the simulation experiments.

Appendix C indicates the technical requirements for consistent estimation of the direct causal effect of particular regressors by TSLS or by KLS in linear regression models with omitted regressors.

A. Findings from the simulation study

The data generating process we used in the Monte Carlo experiments, presented in all its details in Appendix B, is a generalization of those used in the earlier studies on KLS. It concerns a linear regression model for a dependent variable y with an intercept and one slope coefficient β for the single possibly endogenous regressor x . The very simple model $y = c + \beta x + u$ has i.i.d. (independent and identically distributed) disturbances u . The correlation of the regressor and the disturbance, indicated by ρ_{xu} , can be controlled in the experiments. Next to the internal instrument established by the constant, there are two external instrumental variables, z_1 and z_2 . For these their correlation (strength/weakness) with the single regressor can be controlled by ρ_{z_1x} and ρ_{z_2x} respectively. Moreover, their correlation (validity/invalidity) with the disturbance can be controlled by ρ_{z_1u} and ρ_{z_2u} .

For various interesting combinations of ρ_{xu} , ρ_{z_1x} , ρ_{z_2x} , ρ_{z_1u} , ρ_{z_2u} and sample size n we will examine: (i) the rejection probability of the Sargan test at nominal significance level α , where we shall consider $0.01 \leq \alpha \leq 0.5$; and (ii) the estimation errors $\hat{\beta} - \beta$ for various estimators of the slope coefficient, namely OLS, IV (just using the external instrument z_1), TSLS (using both z_1 and z_2) and the new instrument-free estimator KLS. In Appendix B it is proved that all presented findings are invariant with respect to the actual values of the intercept c and slope β , and also to the means of x , z_1 and z_2 . Therefore, without loss of generality, we fixed these all at zero. The results are also invariant regarding the variance of z_1 and of z_2 . Therefore we gave σ_{z_1} and σ_{z_2} value unity. In the graphs below we present the quartiles of the distribution (as assessed from 100,001 replications of the experiments) of the various estimation errors for the case $\sigma_u/\sigma_x = 1$. Outcomes for different σ_u/σ_x ratios can be obtained simply by adapting the scale on the vertical axis accordingly. For the Sargan test we present the rejection frequency over all replications of the simulation for different values of α . These frequencies are in fact not only invariant with respect to β , σ_{z_1} and σ_{z_2} , but also to σ_u and σ_x .

Not all values smaller than one in absolute value for the five correlation coefficients are compatible. For instance, it is self-evidently impossible to have $\rho_{z_1u} = 0$, whereas both ρ_{xu} and ρ_{z_1x} are close to unity. Close to boundary values, and to notoriously problematic cases such as $\rho_{z_1x} \rightarrow 0$, $\rho_{xu} \rightarrow 1$, or n very small, instrumental variable estimators may show pathological behavior. It is not our intention here to demonstrate

that such cases exist and are also problematic for the Sargan test.¹ Our primary aim is here to demonstrate that serious problems occur as well for parameter combinations which at first sight seem pretty harmless. Therefore we start to examine a reasonably large sample ($n = 250$) and rather middle of the road combinations of the correlations arising from:

$$\begin{aligned}\rho_{z_1x} &\in \{0.3, 0.6\}, & \rho_{z_2x} &\in \{0.1, 0.4\}, \\ \rho_{z_1u} &\in \{0.0, 0.1\}, & \rho_{z_2u} &\in \{0.0, 0.2\}, \\ \rho_{xu} &\in \{0.3, 0.6\}.\end{aligned}\tag{A.1}$$

Hence, the instruments will not be chosen ultra-weak, nor extremely invalid. The estimator error quartiles will be examined over the whole range $0 \leq \rho_{xu} \leq 0.9$, but the Sargan test rejection frequency only for the two ρ_{xu} values given in (A.1). The artificial samples drawn are typical for cross-section data, because for all series their n observations are drawn independently. Moreover, we took all of them from the Gaussian distribution.

[Figure A.1 here]

For the various indicated specific situations the four left-hand panels of Figure A.1 contain rejection frequencies of the Sargan test, and the four right-hand panels present the quartiles of the distributions of the four estimation methods compared here. Each row of panels concerns a specific situation regarding instrument (in)validity. Each left-hand panel presents rejection frequencies over a range of nominal significance levels α for the same eight different situations regarding degree of simultaneity and strength of the two instruments. Each right-hand panel presents over a range of values of the endogeneity correlation the three quartiles of the estimation error of the slope coefficient for each of four estimation techniques, and for different situations –when relevant– regarding instrument strength. Therefore, every right-hand panel contains three similarly colored/marketed lines for the same eight different estimators/cases as indicated in the legend. Of course, for each triple of lines the central one is the median. The other two lines give an impression of the dispersion of the distribution of the estimation errors around the median. Their vertical distance represents the interquartile range: of the generated estimation errors 50% landed within these two lines for each ρ_{xu} value.

¹This is one of the main objectives in: Davidson, R., MacKinnon, J.G., 2015. Bootstrap tests for overidentification in linear regression models. *Econometrics* 3, 825-863.

In the top-row of panels both instruments are valid. The top-left panel shows that for the examined eight cases mentioned in the legend the Sargan test shows no noteworthy size problems: the actual probability of type I errors is extremely close to the nominal significance level for all α values examined. In the top-right panel, for all eight estimators/cases represented, except OLS, the three lines are found to be almost horizontal. Thus, these distributions are hardly determined by endogeneity of x , and they suggest median unbiasedness, especially for moderate values of ρ_{xu} . On the other hand, the estimation errors of OLS seem proportional to the degree of endogeneity. Given the relatively small dispersion of OLS and KLS, the graphs show the increasing effects on the dispersion of using weaker and fewer instruments. Note that the dispersion of OLS improves for higher ρ_{xu} and is not beaten by any of the other estimators, although KLS comes close. It is striking that the KLS estimator beats all other estimators when taking both median bias and interquartile range into account. Note, though, that this is the unfeasible version of KLS, which uses full knowledge of the actual value ρ_{xu} . However, one should realize that the instrument based estimators build on assuming ρ_{z_1u} and ρ_{z_2u} both being zero, whereas in practice their true values are in fact unknowable too.

The second graph on the Sargan test shows what the effects are on its rejection probability when one of the two instruments is mildly invalid. When the valid instrument is relatively weak we see that the Sargan test will not very often detect the instrument invalidity. The situation is slightly better when the valid instrument is stronger. However, when using $\alpha = 0.05$ then instrument invalidity will be detected with probability 0.3 at most (for the sample size and correlation combinations examined), so the type II error probability is at least 0.7. The adjacent panel shows that nondetected instrument invalidity (of just $\rho_{z_1u} = 0.1$) is devastating for the estimators based on instruments, especially for the IV estimator just using the invalid instrument. The TSLS estimators based on a valid and an invalid instrument are also seriously affected and for most their interquartile range does no longer overlap with that of KLS. For ρ_{xu} small OLS is in fact to be preferred to IV or TSLS. Note that the OLS and KLS results are similar in all four rows of panels, because they are invariant to the properties of the two instruments.

In the third row of panels instrument z_2 is invalid ($\rho_{z_2u} = 0.2$), so the IV results are similar to those in the top panel. Using $\alpha = 0.5$ would lead for all cases examined to detection of the invalidity with a probability above 0.9, and above 0.5 when using $\alpha = 0.05$. The effect on the estimation errors of TSLS is more determined by the strength of the valid instrument than by the strength of the invalid instrument.

In the fourth row both instruments are invalid, and here we clearly note the perils of the Sargan test not being consistent. For some cases the rejection probability is quite high, but for two of them it hardly exceeds the nominal significance level. These are the two cases where $\rho_{z_1x} = 0.3$ and $\rho_{z_2x} = 0.4$, so the most seriously invalid instrument is also the strongest. The area in the parameter space where the Sargan test will lack power for the chosen data generating process can be derived analytically, see (B.17). The bottom-right panel dramatically undermines trust in instrument-based methods, as this shows that the TSLS estimator for these often Sargan-approved cases is very badly biased over the whole range of ρ_{xu} values. Note that the KLS results are always the most attractive in all four right-hand panels, simply because they are not based on instruments and thus do not require the doubtful approval by the Sargan test.

[Figure A.2 here]

Figure A.2 presents some results for a much smaller and for a much larger sample size than 250. We just cover the cases to be compared with the second and fourth rows of panels of Figure A.1. The size control in this simple cross-section model (not presented in the figure) was found to be close to perfect, irrespective of the sample size (whereas it has been established that size problems for the Sargan/Hansen test are serious in the context of dynamic panel data models). As is to be expected, the detection probability of instrument invalidity is generally lower in smaller samples and larger in bigger samples. For most cases it is (almost) one when $n = 2500$, but even then (and not surprisingly also for $n = 50$) for the same particular cases as in Figure A.1 the detection probability is alarmingly low, with devastating consequences for inference on β as the right-hand graphs show. Note that after a rejection by the Sargan test, producing inference on β requires a further search to find valid instruments.

[Figure A.3 here]

Next we examine the vulnerability of KLS regarding an incorrect assessment r_{xu} of the true endogeneity correlation ρ_{xu} . Figure A.3 presents the quartiles of $\hat{\beta}_{KLS}(r_{xu})$ for all compatible combinations of $\rho_{xu} = -0.9(0.1)0.9$ and $r_{xu} - \rho_{xu} = -0.3(0.1)0.3$, so that $|r_{xu}| < 1$. Results are given for $n = 50, 250, 2500$ and 25000 . The four panels clearly show

that the median of the various distributions of the estimation errors seems invariant to sample size. Apparently the median bias in finite samples is for $n \geq 50$ simply given by the inconsistency of KLS. This inconsistency is only zero when the correct ρ_{xu} value has been used. Of course, the sample size does have a mitigating effect on the interquartile range, which is close to zero when n is very large. We see that, roughly, when $n = 250$ and $|\rho_{xu}| \leq 0.4$ an error of ± 0.3 may give rise to a shift of the quartiles of up to about $0.5 \times \sigma_u/\sigma_x$, and about half of that for errors of ± 0.2 . The latter vulnerability, although substantial, seems more limited than that of IV/TSLS when using mildly weak and/or mildly invalid instruments.

Figure A.1 Simulation results for $n = 250$; $\sigma_x/\sigma_u = 1$; and correlations (A.1)

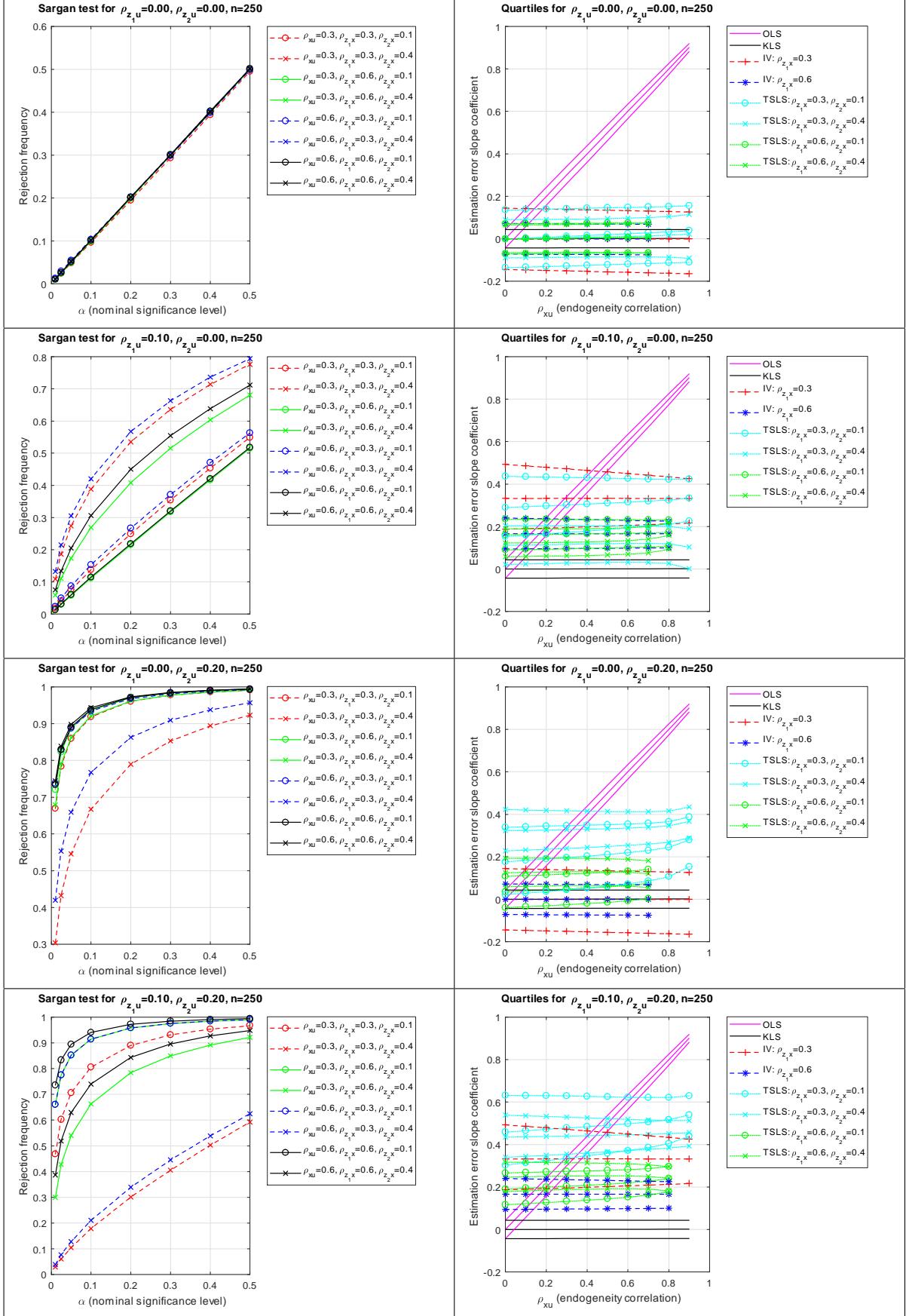


Figure A.2 Simulation results for $n = 50, 2500$; $\sigma_x/\sigma_u = 1$; and correlations (A.1)

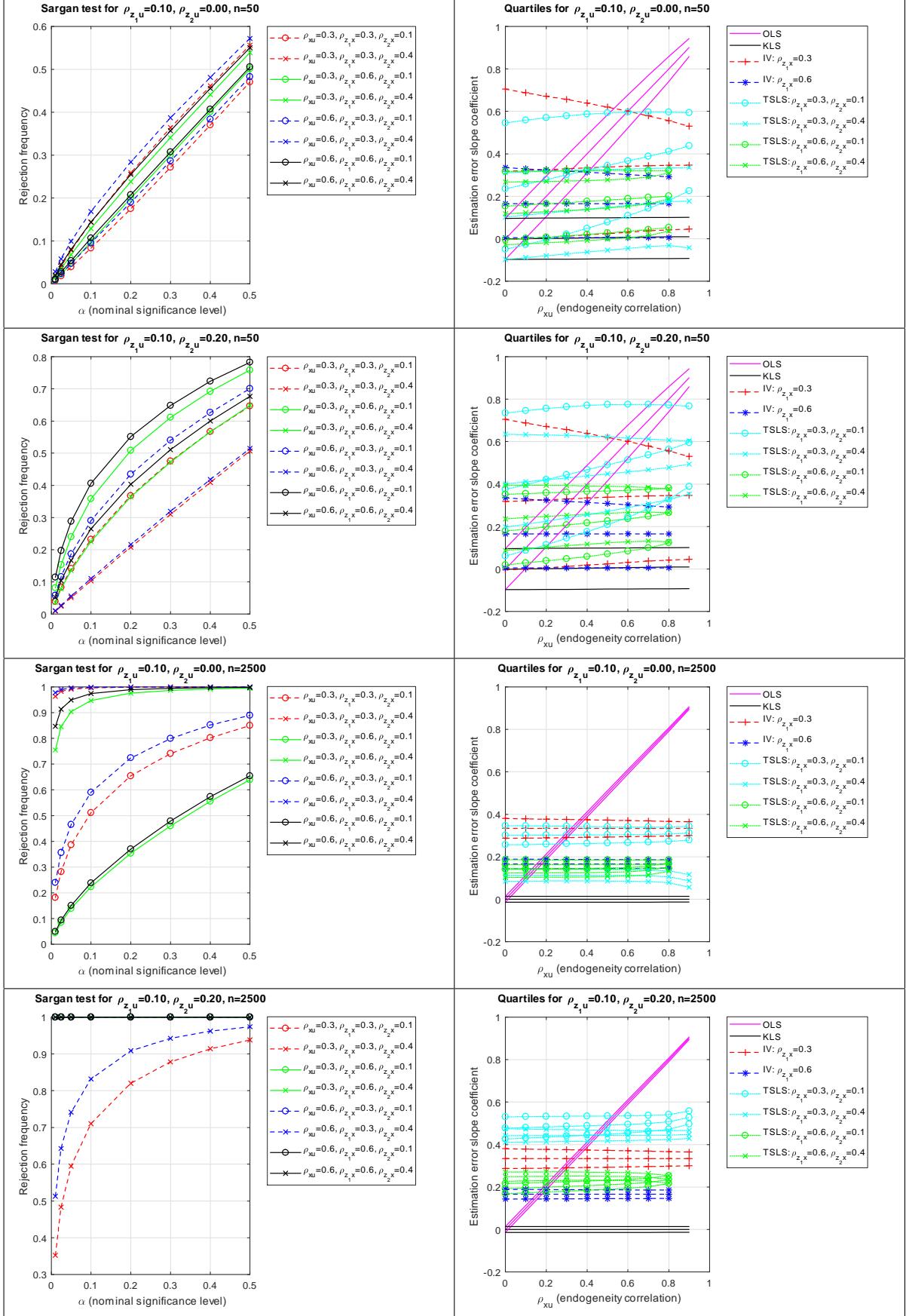
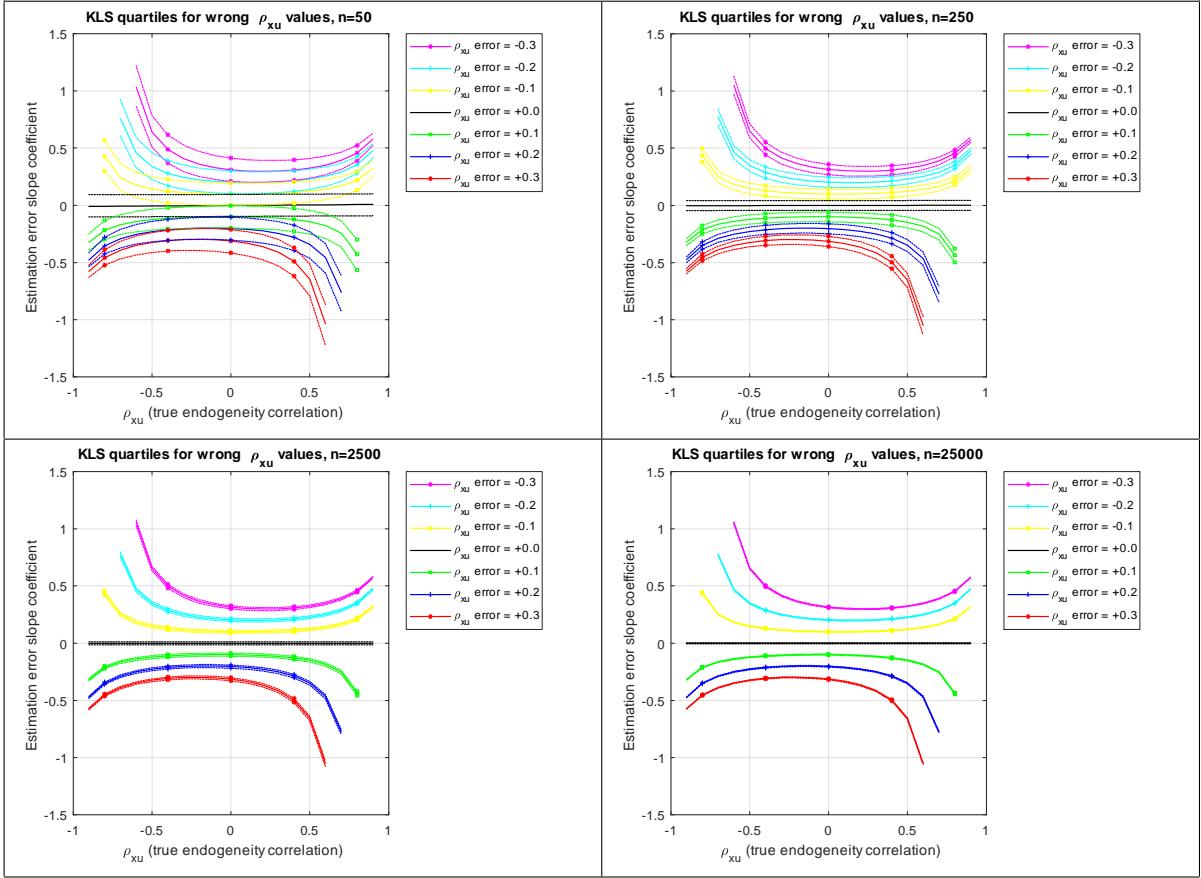


Figure A.3 KLS quantiles using a wrong ρ_{xu} for $n = 50, 250, 2500, 25000$.



B. Further details on the chosen simulation design

The Monte Carlo design used in Appendix A is defined as follows. Let ε_i , ξ_i , ζ_{i1} and ζ_{i2} be four mutually independent series ($i = 1, \dots, n$) of identically distributed independent drawings with mean zero and unit variance. From these we construct the four series

$$u_i = \sigma_u \varepsilon_i \sim iid(0, \sigma_u^2), \quad (\text{B.1})$$

$$x_i = \sigma_x [(1 - \rho_{xu}^2)^{1/2} \xi_i + \rho_{xu} \varepsilon_i] \sim iid(0, \sigma_x^2), \quad (\text{B.2})$$

$$z_{ij} = \sigma_{z_j} (\rho_{z_j \zeta_j} \zeta_{ji} + \rho_{z_j \xi} \xi_i + \rho_{z_j u} \varepsilon_i) \sim iid(0, \sigma_{z_j}^2) \text{ for } j = 1, 2, \quad (\text{B.3})$$

where all ρ coefficients do not exceed 1 in absolute value; moreover,

$$\rho_{z_j \zeta_j}^2 + \rho_{z_j \xi}^2 + \rho_{z_j u}^2 = 1 \text{ for } j = 1, 2. \quad (\text{B.4})$$

Obviously, $\sigma_{xu} = \rho_{xu} \sigma_x \sigma_u$, $\sigma_{z_j u} = \rho_{z_j u} \sigma_{z_j} \sigma_u$ and $\sigma_{z_j x} = \sigma_{z_j} \sigma_x [\rho_{z_j \xi} (1 - \rho_{xu}^2)^{1/2} + \rho_{z_j u} \rho_{xu}]$, hence $\rho_{z_j x} = \rho_{z_j \xi} (1 - \rho_{xu}^2)^{1/2} + \rho_{z_j u} \rho_{xu}$, which yields

$$\rho_{z_j \xi} = (\rho_{z_j x} - \rho_{z_j u} \rho_{xu}) (1 - \rho_{xu}^2)^{-1/2}, \quad (\text{B.5})$$

for $\rho_{xu}^2 < 1$. From (B.4) we also have

$$\rho_{z_j \zeta_j} = (1 - \rho_{z_j \xi}^2 - \rho_{z_j u}^2)^{1/2}. \quad (\text{B.6})$$

Hence, when values for $\sigma_u > 0$, $\sigma_x > 0$, $\sigma_{z_j} > 0$, $|\rho_{xu}| < 1$, $|\rho_{z_j x}| \leq 1$ and $|\rho_{z_j u}| \leq 1$ are chosen, we can generate series for u_i and x_i and find from (B.5) matching values for $\rho_{z_j \xi}$ and for $\rho_{z_j \zeta}$ from (B.6), so that series z_{i1} and z_{i2} can be generated as well. However, the choices for ρ_{xu} , $\rho_{z_j x}$ and $\rho_{z_j u}$ are only compatible if they yield values for $\rho_{z_j \xi}^2$ and $\rho_{z_j \zeta_j}^2$ such that $0 \leq \rho_{z_j \xi}^2 \leq 1$ and $0 \leq \rho_{z_j \zeta_j}^2 \leq 1$. This requires

$$0 \leq \frac{(\rho_{z_j x} - \rho_{z_j u} \rho_{xu})^2}{(1 - \rho_{xu}^2)} \leq 1 \quad (\text{B.7})$$

and

$$0 \leq 1 - \frac{(\rho_{z_j x} - \rho_{z_j u} \rho_{xu})^2}{(1 - \rho_{xu}^2)} - \rho_{z_j u}^2 \leq 1. \quad (\text{B.8})$$

The latter implies

$$0 \leq 1 - \rho_{xu}^2 - (\rho_{z_j x} - \rho_{z_j u} \rho_{xu})^2 - \rho_{z_j u}^2 (1 - \rho_{xu}^2) \leq 1 - \rho_{xu}^2$$

or

$$0 \leq (1 - \rho_{xu}^2) (1 - \rho_{z_j u}^2) - (\rho_{z_j x} - \rho_{z_j u} \rho_{xu})^2 \leq 1 - \rho_{xu}^2,$$

giving the two requirements

$$(\rho_{z_jx} - \rho_{z_ju}\rho_{xu})^2 \leq (1 - \rho_{xu}^2)(1 - \rho_{z_ju}^2), \quad (\text{B.9})$$

and

$$-\rho_{z_ju}^2(1 - \rho_{xu}^2) - (\rho_{z_jx} - \rho_{z_ju}\rho_{xu})^2 \leq 0.$$

The latter will always be satisfied, whereas restriction (B.9) implies that (B.7) will also be satisfied.

So, by choosing values $|\rho_{xu}| < 1$, $|\rho_{z_jx}| \leq 1$ and $|\rho_{z_ju}| \leq 1$, which obey (B.9), we have two instruments with correlation

$$\begin{aligned} \rho_{z_1z_2} &= \rho_{z_1\xi}\rho_{z_2\xi} + \rho_{z_1u}\rho_{z_2u} \\ &= (\rho_{z_1x} - \rho_{z_1u}\rho_{xu})(\rho_{z_2x} - \rho_{z_2u}\rho_{xu})(1 - \rho_{xu}^2)^{-1} + \rho_{z_1u}\rho_{z_2u}. \end{aligned}$$

For each realization of the series u_i , x_i and z_{ij} in the simulation replications, we may first subtract their respective sample average from each observation. In that way we cover a model with one slope coefficient and an arbitrary intercept, to be estimated by OLS, KLS, IV or TSLS, because there are next to the intercept two (possibly invalid) instruments, each with a possibly non-zero arbitrary mean which has been partialled out. The dependent variable is generated by the model

$$y_i = x_i\beta + u_i. \quad (\text{B.10})$$

This can be estimated by

$$\begin{aligned} \hat{\beta}_{OLS} &= \frac{x'y}{x'x} = \beta + \frac{x'u}{x'x}, \\ \hat{\beta}_{KLS} &= \hat{\beta}_{OLS} - \rho_{xu} \left(\frac{\hat{u}'_{OLS}\hat{u}_{OLS}}{x'x} \right)^{1/2}, \text{ where } \hat{u}_{OLS} = y - x\hat{\beta}_{OLS}, \hat{u}'_{OLS}\hat{u}_{OLS} = u'u - \frac{(u'x)^2}{x'x}, \\ \hat{\beta}_{IV}^{(j)} &= \frac{z'_jy}{z'_jx} = \beta + \frac{z'_ju}{z'_jx}, \quad j = 1, 2, \\ \hat{\beta}_{TSLS} &= \frac{x'Z(Z'Z)^{-1}Z'y}{x'Z(Z'Z)^{-1}Z'x} = \beta + \frac{x'P_Zu}{x'P_Zx}, \text{ with } Z = (z_1, z_2) \text{ and } P_Z = Z(Z'Z)^{-1}Z'. \end{aligned}$$

For the estimation errors we find, writing $\xi_i^* = (1 - \rho_{xu}^2)^{1/2}\xi_i + \rho_{xu}\varepsilon_i$,

$$\hat{\beta}_{OLS} - \beta = \frac{\sigma_u \Sigma_i \xi_i^* \varepsilon_i}{\sigma_x \Sigma_i \xi_i^{*2}}, \quad (\text{B.11})$$

$$\begin{aligned} \hat{\beta}_{KLS} - \beta &= \frac{\sigma_u \Sigma_i \xi_i^* \varepsilon_i}{\sigma_x \Sigma_i \xi_i^{*2}} - \rho_{xu} \left\{ \frac{\sigma_u^2 \Sigma_i \varepsilon_i^2}{\Sigma_i x_i^2} - \frac{\sigma_u^2 [\Sigma_i \varepsilon_i x_i]^2}{[\Sigma_i x_i^2]^2} \right\}^{1/2} \\ &= \frac{\sigma_u}{\sigma_x} \left\{ \frac{\Sigma_i \xi_i^* \varepsilon_i}{\Sigma_i \xi_i^{*2}} - \rho_{xu} \left[\frac{\Sigma_i \varepsilon_i^2}{\Sigma_i \xi_i^{*2}} - \frac{\Sigma_i \xi_i^* \varepsilon_i}{(\Sigma_i \xi_i^{*2})^2} \right]^{1/2} \right\}, \end{aligned} \quad (\text{B.12})$$

$$\hat{\beta}_{IV}^{(j)} - \beta = \frac{\sigma_u}{\sigma_x} \frac{\Sigma_i (\rho_{z_j \zeta} \zeta_{ij} + \rho_{z_j \xi} \xi_i + \rho_{z_j u} \varepsilon_i) \varepsilon_i}{\Sigma_i (\rho_{z_j \zeta} \zeta_{ij} + \rho_{z_j \xi} \xi_i + \rho_{z_j u} \varepsilon_i) \xi_i^*} \quad (\text{B.13})$$

$$\hat{\beta}_{TSLS} - \beta = \frac{x' P_Z u}{x' P_Z x}. \quad (\text{B.14})$$

Because P_Z is invariant with respect to the scale of the vectors $z^{(1)}$ and $z^{(2)}$ the estimation errors of $TSLS$, like those of IV are invariant with respect to σ_{z_1} and σ_{z_2} , so without loss of generality we may fix these at value 1.

It is easily seen that all the estimation errors are also invariant regarding β and are all a multiple of σ_u/σ_x . Hence, without loss of generality we may choose in the simulations $\beta = 0$, $\sigma_{z_1} = \sigma_{z_2} = 1$ and $\sigma_x = 1$. Then the dispersion of all estimators can be regulated by varying σ_u . However, their **relative** differences will be invariant with respect to σ_u . So, by just choosing $\sigma_u = 1$ all relevant information will be obtained, through choosing relevant compatible values for the remaining design parameters: n , ρ_{xu} , $\rho_{z_j x}$ and $\rho_{z_j u}$, where the latter two determine $\rho_{z_j \xi}$ and $\rho_{z_j \zeta_j}$. Changing the sign of any of the correlations while keeping their absolute value fixed has simple (anti-)symmetric effects just on the sign of the estimation errors. Therefore we shall mostly just investigate nonnegative values for ρ_{xu} , $\rho_{z_j x}$ and $\rho_{z_j u}$.

For the TSLS residuals we find

$$\hat{u}_{TSLS} = y - \hat{\beta}_{TSLS} x = u - \frac{x' Z (Z' Z)^{-1} Z' u}{x' Z (Z' Z)^{-1} Z' x} x = u - \frac{x' P_Z u}{x' P_Z x} x, \quad (\text{B.15})$$

and for the Sargan test statistic

$$\begin{aligned} S &= n \frac{\hat{u}'_{TSLS} Z (Z' Z)^{-1} Z' \hat{u}_{TSLS}}{\hat{u}'_{TSLS} \hat{u}_{TSLS}} \\ &= n \frac{u' P_Z u - \frac{(x' P_Z u)^2}{x' P_Z x}}{u' u - 2u' x \frac{x' P_Z u}{x' P_Z x} + x' x \frac{(x' P_Z u)^2}{(x' P_Z x)^2}} \\ &= n \frac{u' P_Z u (x' P_Z x)^2 - (x' P_Z u)^2 x' P_Z x}{u' u (x' P_Z x)^2 - 2u' x (x' P_Z u) x' P_Z x + x' x (x' P_Z u)^2}. \end{aligned} \quad (\text{B.16})$$

It is obvious that this is invariant with respect to β and to all scale factors, because all individual terms, both in the numerator and in the denominator, are multiples of $\sigma_u^2 \sigma_x^4$.

It is well known that the Sargan test is equivalent to literally testing over-identification exclusion restrictions. In the present design, this amounts to estimating the model $y_i = \beta x_i + \delta z_{ij} + u_i$, where j is either 1 or 2, using both instruments, and next testing the significance of δ . One easily finds that the probability limit of the estimator for δ is a multiple of $\rho_{z_1x}\rho_{z_2u} - \rho_{z_2x}\rho_{z_1u}$, which is zero when

$$\rho_{z_1u}/\rho_{z_1x} = \rho_{z_2u}/\rho_{z_2x}. \quad (\text{B.17})$$

Indeed, when running simulations with parameter values obeying (B.17) with values of ρ_{z_1u} and ρ_{z_2u} far away from zero, we always found rejection probabilities of the Sargan test similar to the nominal significance level, with TSLS inference of course being seriously corrupted.

C. Coping with omitted variables by TSLS or KLS

We consider estimating the model

$$y = X\beta + u, \quad (\text{C.1})$$

as introduced in (4.1) and (4.2), where $X\beta = X_1\beta_1 + X_2\beta_2$, $u = \gamma + \varepsilon$, $\gamma = X_3\beta_3$ and $E(\varepsilon | X_1, X_2, X_3) = 0$. For an $n \times L$ instrument matrix Z and $n \times K$ matrix X , obeying the conditions

$$\text{rank}(X) = K_1 + K_2 = K, \text{ rank}(Z) = L \text{ and } \text{rank}(Z'X) = K \leq L, \quad (\text{C.2})$$

the TSLS estimator

$$\hat{\beta}_{TSLS} = (X'P_Z X)^{-1} X' P_Z y, \text{ where } P_Z = Z(Z'Z)^{-1}Z', \quad (\text{C.3})$$

exists. Under some further standard regularity conditions, including the moment conditions

$$E[Z'(\gamma + \varepsilon)] = 0, \quad (\text{C.4})$$

$\hat{\beta}_{TSLS}$ is consistent for $\beta = (\beta_1, \beta_2)'$.

Let X_2 contain all the regressors of (C.1) that are being used as internal instruments. Then $Z = (Z_1, X_2)$, with Z_1 containing $L - K_2$ external instruments, which are not in the space spanned by the columns of X . Now (C.2) requires

$$\text{rank}(Z_1'X_1) = K_1, \quad (\text{C.5})$$

and (C.4) specializes to $E(Z_1'\gamma) + E(Z_1'\varepsilon) = 0$ and $E(X_2'\gamma) + E(X_2'\varepsilon) = 0$. Using $E(X_2'\varepsilon) = 0$, these moment conditions boil down to

$$E(Z_1'\gamma) + E(Z_1'\varepsilon) = 0 \text{ and } E(X_2'\gamma) = 0. \quad (\text{C.6})$$

In theory the first condition of (C.6) could be satisfied by finding external instruments Z_1 such that $E(Z_1'\varepsilon) = -E(Z_1'\gamma) \neq 0$. However, realizing this in practice seems elusive, whereas achieving (C.6) by pursuing validity of the sufficient conditions

$$(i) E(Z_1'\varepsilon) = 0, (ii) E(Z_1'\gamma) = 0, \text{ and (iii)} E(X_2'\gamma) = 0 \quad (\text{C.7})$$

may be feasible. Here condition (i) entails that the variables Z_1 should be correctly excluded from the theory model (4.1). Hence, external instruments Z_1 should be chosen

such that they do not have a direct effect on y , but just an indirect effect (if $\beta_1 \neq 0$) through their required association with X_1 , imposed by (C.5), which prevents under-identification. Assuming (i) holds, condition (ii) is explicitly addressed by arguing or testing that Z_1 establishes valid external instruments for the empirical model (C.1), requiring the variables in Z_1 to be uncorrelated with ε and γ . Condition (iii) of (C.7) is satisfied if the variables X_2 , which are uncorrelated with ε , are uncorrelated with γ too, and thus are (like Z_1) uncorrelated with the disturbances $\gamma + \varepsilon$ of the empirical model (C.1). Whether or not the instruments Z_1 are correlated with the regressors X_2 is irrelevant for the consistency of TSLS using instruments (Z_1, X_2) , as long as such correlations are not extreme and would jeopardize (C.2).

Conditions (ii) and (iii) may often be hard to fulfill. In the applications of Section 4, Z_1 may contain a variable such as presence of a college in the neighborhood or quarter of birth. If these may be correlated with ability, for instance, because both Z_1 and γ depend on variables from X_2 , such as urban or race, then both requirements (ii) and (iii) are at risk. However, such doubts to ascertain moment conditions (ii) and (iii) can be mitigated if one is willing to focus on obtaining a consistent estimator for just the direct effects β_1 of X_1 , and give that up regarding β_2 , while supposing that sufficiently relevant control variables X_2 have been included in the model.

When both (ii) and (iii) are not fulfilled simply because both γ and Z_1 are correlated with X_2 , then the situation can be characterized and tackled as follows. We define

$$\gamma^* = \gamma - X_2\phi_2^*, \text{ where } E(\gamma | X_2) = X_2\phi_2^*, \quad (\text{C.8})$$

$$Z_1^* = Z_1 - X_2\Psi, \text{ where } E(Z_1 | X_2) = X_2\Psi, \quad (\text{C.9})$$

and suppose

$$E(Z_1^{*\prime}\gamma^*) = 0. \quad (\text{C.10})$$

Now model (C.1) can be rewritten as

$$y = X_1\beta_1 + X_2\beta_2^* + (\gamma^* + \varepsilon), \text{ where } \beta_2^* = \beta_2 + \phi_2^*. \quad (\text{C.11})$$

Supposing (i) and (C.10) are fulfilled yields the orthogonality conditions

$$E[Z_1'(\gamma^* + \varepsilon)] = E(Z_1^*\gamma^*) + \Psi'E(X_2'\gamma^*) = \Psi'E[X_2'E(\gamma^* | X_2)] = 0, \quad (\text{C.12})$$

$$E[X_2'(\gamma^* + \varepsilon)] = E\{E[X_2'(\gamma - X_2\phi_2^*) | X_2]\} = E(X_2'X_2\phi_2^* - X_2'X_2\phi_2^*) = 0, \quad (\text{C.13})$$

so that estimating model (C.11) by TSLS using the instruments (Z_1, X_2) delivers consistent estimators of β_1 and β_2^* .

The above analysis highlights that as a rule instrument matrix Z should not contain the variables from X for which one wants to assess exclusively their direct effect on y in isolation from any indirect effects they may have on y through omitted regressors. These variables are collected in X_1 . The remaining regressors from X , the variables X_2 , can be used as internal instruments, but their estimated coefficients will represent both their direct effect and any indirect effect they may have via the omitted variables X_3 . For identification of β_1 and β_2^* one needs at least as many external instruments Z_1 as there are variables in X_1 . These external instruments Z_1 should meet three criteria. Two with respect to their validity, which can be formulated as: (a) Z_1 should have no direct causal effect on y ; and (b), after removing from Z_1 and from omitted component γ any linear association with X_2 , they should be mutually uncorrelated. In addition, Z_1 should meet a third criterion, namely regarding relevance, being: (c) matrix $Z_1'(I - P_{X_2})X_1$ should be such that the external instruments Z_1 are sufficiently strong, in the sense that the correlations of the variables in Z_1 with those of X_1 , after netting out their relation with X_2 , should not be small.

From the above it follows that an investigator may choose how many and which of the regressors (s)he is willing to consider as exogenous and let slip the ambition to consistently estimate their direct effect. Provided valid external instruments will be found the direct effect of the remaining endogenous regressors can be estimated consistently. In principle, one may even choose to estimate the direct effect of various of the regressors X one by one, treating each of them as the one and only endogenous regressor in a series of regressions.

Likewise, in the instrument-free approach, one may choose to analyze a series of models in which the regressor matrix X of model (C.1) is partitioned in a different way into endogenous and exogenous regressors. This partition of X determines which of the elements of ρ_{xu} can simply be set at zero, and for which credible numerical (interval) assumptions have to be made. Of course, the other crucial issues in a TSLS context, indicated in the previous paragraph by (a), (b) and (c), are irrelevant for an instrument-free approach. When avoiding assumptions on instruments, these have to be replaced by an assumption on credible numerical values of $\rho_{x^{(1)}u}$, which is associated with the covariance $E(X_1'u) = E(X_1'\gamma^*)$. The j^{th} element of $\rho_{x^{(1)}u}$ can be obtained by dividing the corresponding element of $E(X_1'u)$ by $n\sigma_u\sigma_j$, where σ_j is the standard deviation of the j^{th} regressor in X_1 .