



IZA DP No. 3961

On the Sensitivity of Return to Schooling Estimates to Estimation Methods, Model Specification, and Influential Outliers If Identification Is Weak

David A. Jaeger
Juliane Parys

January 2009

On the Sensitivity of Return to Schooling Estimates to Estimation Methods, Model Specification, and Influential Outliers If Identification Is Weak

David A. Jaeger

*City University of New York
and IZA*

Juliane Parys

University of Bonn

Discussion Paper No. 3961
January 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ABSTRACT

On the Sensitivity of Return to Schooling Estimates to Estimation Methods, Model Specification, and Influential Outliers If Identification Is Weak^{*}

We provide a comparison of return to schooling estimates based on an influential study by Angrist and Krueger (1991) using two stage least squares (TSLS), limited information maximum likelihood (LIML), jackknife (JIVE), and split sample instrumental variables (SSIV) estimation. We find that the estimated return to education is quite sensitive to the age controls used in the models as well as the estimation method used. In particular, we provide evidence that JIVE coefficients' standard errors are inflated by a group of extreme years of education observations, for which identification is especially weak. We propose to use Cook's Distance in order to identify influential outliers having substantial influence on first-stage JIVE coefficients and fitted values.

JEL Classification: C13, C31, J31

Keywords: Cook's Distance, heteroskedasticity, outliers, return to education, specification, weak instruments

Corresponding author:

David A. Jaeger
The Graduate Center
City University of New York
365 Fifth Avenue
New York, NY 10016
USA
E-mail: djaeger@gc.cuny.edu

^{*} The authors wish to thank Joerg Breitung and Guido Imbens for very helpful discussions and comments. All errors are our own.

1 Introduction

A recent and large literature has examined the quality of inference based on traditional two stage least squares (TSLS) or limited information maximum likelihood (LIML) methods when the correlation between the endogenous regressors and the instruments is weak. This literature has attempted to develop better estimators and more credible inference methods. Although the inappropriateness of conventional normal approximations in the presence of weak or no identification was discussed in earlier work, an empirical study by Angrist and Krueger (1991) (henceforth, AK) and the critiques by Bound et al. (1995) (henceforth BJB) and Staiger and Stock (1997) have mainly motivated the recent weak instruments literature, where conventional large sample approximations were *ex post* detected to be misleading. The AK results are an important benchmark in much of the recent micro-econometric literature on weak instruments (see Stock et al. (2002) for an overview of the issues).

One of the lessons from the BJB's subsequent analysis of AK's data is that interactions of exogenous covariates with one or a few basic instruments in an attempt to improve precision can lead to substantially biased TSLS results, even with a high degree of overidentification and hundreds of thousands of observations. In such cases, conventional inference methods can be misleading. Although the LIML estimator usually performs better than TSLS (in terms of bias), conventional LIML standard errors will be too small. In this paper, we discuss two relatively new estimators aiming to overcome the weaknesses of the TSLS and LIML estimators. Angrist and Krueger (1995) (henceforth AK-SSIV) propose a split sample instrumental variables (SSIV) estimator that requires an arbitrary sample split while a jackknife instrumental variable estimator (JIVE) is proposed by Angrist et al. (1999) (henceforth AIK-JIVE) and uses a leave-one-out jackknife-type fitted value instead of the first-stage equation of TSLS.

We examine the performance of all of these estimators with AK's data, paying special attention to specification differences and the strength of instruments among different groups of observations. We find that all of the estimators suffer from at least one defect. JIVE seems to perform poorer than other IV estimators when some groups of observations have substantial influence on first-stage coefficients and hence predicted values. LIML suffers from a potential heteroskedasticity problem in the errors that may lead them to be inconsistent. The estimates from SSIV are quite imprecise and depend very much on the random sample split and the data set used. We find that none of the estimators is able to provide precisely estimated and believable estimates for the return to education when controlling for important wage determining variables available in the U.S Census, particularly age, and when using the

full sample.

The rest of the paper is organized as follows. In section 2 we briefly present AK's identification strategy and discuss *ex post* problems that we find with it. In section 3 we discuss general problems with conventional estimators and weak identification as well as briefly present the JIVE and SSIV estimators that address those issues. In section 4 we report our recalculations of return to education coefficients and in section 5 we briefly describe the implications of heteroskedasticity for the results. In section 6 we discuss the problem of outliers in the first-stage regression. We offer some conclusions in Section 7.

2 Angrist and Krueger's Setup

AK was a landmark in using instrumental variables to estimate the returns to education (Card (1999)) and uses, perhaps, the most credible identification strategy of all papers in the literature. They consider the following model:

$$y = X\beta + u , \tag{1}$$

where y is log weekly earnings, and X contains a constant, years of education, variables controlling for age, and in some specifications further relevant covariates. Following an extensive literature in labor economics, AK treat years of schooling as an endogenous variable, while all other variables are treated as exogenous. Education is not assumed to be randomly assigned, but may be systematically related to the unobserved determinants of wages, u , inducing a bias in ordinary least squares estimates.

To address this endogeneity issue, AK exploit the association between quarter of birth and years of education arising from compulsory school attendance and school starting laws. During the cohorts that AK study, the typical U.S. school district required students to have turned six in the calendar year in which they entered school and to stay in school until their sixteenth birthday. Thus, individuals born in the fourth quarter may obtain higher levels of schooling than those born earlier in the year. AK also attempt to exploit variation across states and over time in compulsory schooling laws to identify the returns to education. The crux of their argument is, however, that the interaction of school age starting laws and compulsory school attendance laws lead to variation in years of schooling by quarter of birth, and that this variation is exogenous to the determination of wages. For TSLS estimation, the first-stage equation is

then given by

$$X = Z\pi + v . \tag{2}$$

Z is the matrix of instruments containing quarter of birth dummies interacted with year of birth dummies and the exogenous variables of X . π is the corresponding vector of coefficients.

AK use data from the five percent Public Use Microsample of the 1980 U.S. Census for cohorts of men being born during the 1920s, 1930s and 1940s. As with much of the subsequent literature that uses AKs data to explore the consequences of weak instruments, we limit our attention to the cohort of men born between 1930 and 1939. The identifying variation in the first stage comes from the relationship between quarter of birth and education. But because wages also vary with labor market experience, and hence closely with age, controlling for age is also important, even in the relatively older cohorts that AK study. As there are four quarters of birth in each of the 10 birth-year cohorts that we examine, there are 40 degrees of freedom associated with age in total. AK report estimated coefficients that are highly dependent on the estimator and specification used. Besides a constant and years of schooling, X contains the following variables:

- *Specification (1)*: 9 year of birth dummies
- *Specification (2)*: 9 year of birth dummies, race (1=black), smsa (1=central city), marital status (1=married, living with spouse), and 8 regional dummies
- *Specification (3)*: age (measured in quarters), age squared, 9 year of birth dummies
- *Specification (4)*: age, age squared, 9 year of birth dummies, race, smsa, marital status, and 8 regional dummies

In specifications 1 and 2 identification comes from within birth-year differences in schooling. When controlling for age and age squared, however, identification comes from deviations from the smooth age function, allowing for changes across year of birth cohorts. By adding age and age squared we control for within year of birth age effects on earnings. As it was pointed out by AK and van der Klaauw (2002), the inclusion of age and age squared leads to a considerable change in the estimated return to one additional year of education. Age and age squared seem to be important omitted variables in the first two specifications. Therefore, following BJB, we consider specification 4 to be the preferred one.

BJB were the first to show that because of the weak correlation between years of education and the instruments used, in some specifications the normal approximations used by AK may

not be appropriate, despite the large sample size of almost 330,000 observations. Especially in specifications 3 and 4 the instruments explain little of the variation in schooling as age and age squared already use up lots of the variation within the 40 birth year \times quarter of birth degrees of freedom. The most damning evidence presented by BJB are those where they recalculate the AK estimates using randomly-generated rather than actual quarter of birth to create the instruments and obtain quite reasonable looking estimates. The recent weak instrument literature was largely motivated by these results. As noted by Cruz and Moreira (2005) and others, however, the usual Gaussian large sample approximations for TSLS and LIML are not valid in this case, as they rely on non-zero correlations between the endogenous regressor and the instruments. If the instruments' coefficients are close to zero, the TSLS and LIML estimator may not be close to normally distributed.

Staiger and Stock (1997) formalize the problem. They consider an alternative asymptotic sequence and model the instruments' coefficient matrix, π , as a function of the sample size, n , such that $\pi = c/\sqrt{n}$. In that specific sequence, the concentration parameter $\pi'Z'Z\pi/\sigma_v^2$ converges to a constant, where σ_v^2 is the error variance of the first-stage equation. The TSLS estimator is not consistent in that case but converges to a nondegenerate distribution similar to a multivariate t distribution. Under this alternative asymptotic approximation Staiger and Stock (1997) compare coverage properties of different confidence intervals. Their main concern is that if under the alternative asymptotics a particular confidence interval does not have the correct coverage, then in a potentially important part of the parameter space around $\pi_i = 0$ there are values of the error parameters so that the nominal coverage is considerably away from the actual coverage independent of the sample size. More recently, Mikusheva (2007) therefore formulated the requirement of asymptotically correct coverage probabilities of confidence intervals over the whole parameter space.

3 Alternative Estimators

As an alternative to the TSLS and LIML estimators, two estimators have been proposed that are based on the idea of estimating the first-stage equation on only part of the data. The fundamental idea in both is to break any correlation between the errors in the structural and reduced form equations. In the first of these, AK-SSIV proposed a split sample instrumental variable (SSIV) estimator that is not as biased toward OLS as TSLS. In SSIV procedure, parameters of the first stage are estimated using one randomly chosen half of the sample.

These first-stage estimates are then used to calculate the fitted values for the other half of the sample, which is used to obtain the second-stage parameters. The estimator is defined as

$$\hat{\beta}_{\text{SSIV}} = [X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Z_1(Z_2'Z_2)^{-1}Z_2'X_2]^{-1}X_2'Z_2(Z_2'Z_2)^{-1}Z_1'y_1, \quad (3)$$

where the indices denote the sample half. In the above notation the first-stage equation is estimated with sample half 2. AK-SSIV argue that the SSIV estimator is biased toward zero. They therefore propose a modified estimator that is unbiased under so-called “group asymptotics”, where each cross-section replication provides m additional observations. In the AK data this could mean obtaining additional instruments by adding new cross-sections for new years of data, from new states, regions, or cohorts. The unbiased split-sample instrumental variables (USSIV) estimator is given by

$$\hat{\beta}_{\text{USSIV}} = [X_2'Z_2(Z_2'Z_2)^{-1}Z_1'X_1]^{-1}X_2'Z_2(Z_2'Z_2)^{-1}Z_1'y_1. \quad (4)$$

One potential reason why the SSIV and USSIV estimator has not been explored extensively in the literature is that that, at least with weak instruments, the point estimates are highly dependent on the random sample split.

Subsequent to AK-SSIV, a different alternative estimator was proposed by Angrist et al. (1999). As with SSIV, the idea behind the jackknife instrumental variables estimator (JIVE) is to construct an instrument that is independent of the error in the reduced form equation, especially in small samples. A “leave-one-out” jackknife-type fitted value is used to eliminate the correlation between the estimate of $Z_i\pi$ and X_i . This estimator is first-order equivalent to TSLS, but it has better finite sample properties. The JIVE1 estimator is given by

$$\hat{\beta}_{\text{J1}} = [\hat{X}'_{\text{J1}}X]^{-1}\hat{X}'_{\text{J1}}y, \quad (5)$$

where \hat{X}_{J1} is a n by k matrix with i^{th} row

$$Z_i(Z(i)'Z(i))^{-1}Z(i)'X(i) = Z_i \frac{(Z'Z)^{-1}}{1 - Z_i(Z'Z)^{-1}Z_i'}(Z'X - Z_i'X_i). \quad (6)$$

In this notation, $X(i)$ and $Z(i)$ are the matrices X and Z , respectively, without the i^{th} row. k is the number of instruments.

The JIVE2 estimator adjusts only the $Z'X$ component of $\hat{\pi} = (Z'Z)^{-1}Z'X$ to achieve

unbiasedness. The JIVE2 estimator is defined as

$$\hat{\beta}_{J2} = [\hat{X}'_{J2}X]^{-1}\hat{X}'_{J2}y, \quad (7)$$

with

$$Z_i(Z'Z)^{-1}Z(i)'X(i)\frac{N}{N-1} = Z_i\frac{(Z'Z)^{-1}}{1-1/N}(Z'X - z'_iX_i) \quad (8)$$

being the i^{th} row of \hat{X}_{J2} . Chao et al. (2007) show the asymptotic normality of JIVE estimators. They consider the many instruments sequence of Bekker (1994) and the many weak instruments sequence of Chao and Swanson (2005) and show the consistency of JIVE standard errors under much weaker conditions on the strength of instruments than TSLS and LIML standard errors require to be consistent even in the presence of heteroskedasticity. Of the estimators discussed here, it would appear that JIVE has the most advantageous properties under weak identification. This holds especially if the errors are heteroskedastic.

4 Estimation results

The estimated return to education from AK, AK-SSIV and AIK-JIVE for the cohort of 329,509 men born 1930-1939 are reported in Table 1. For TSLS, JIVE and SSIV estimation 30 instruments are used, which are generated by interacting quarter of birth with year of birth dummies. As the first-stage F statistic and the partial R^2 are relatively small, quarter of birth \times year of birth interactions seem to have very poor explanatory power as instruments for the endogenous years of schooling variable. As proposed by BJB and later improved by Stock and Yogo (2002), the first-stage F statistic on the excluded instruments serves as an indicator in order to detect weak identification. Given the systematic relationship between the instruments and age measured in quarters, it is not surprising that the first-stage F statistic becomes even smaller in those specifications controlling for age and age squared.

Although AK find significant and fairly plausible-looking TSLS coefficients for the return to education in all specifications, BJB and AK-SSIV point out the potentially substantial bias of TSLS coefficients towards OLS when the instruments are very weakly correlated with the endogenous explanatory variable. In reaction to BJB's critique, AIK-JIVE present some results that the TSLS, LIML, and JIVE estimators give similar estimates on the return to education. They interpret this finding as evidence that the TSLS coefficients reported in AK are not severely biased. In addition to the SSIV coefficients of specification (3) reported in AK-SSIV

and the LIML and JIVE coefficients for specification (1) given in AIK-JIVE, in Table 1 we also provide coefficients from all estimators in all four specifications. The variation in the results makes clear that the specification chosen has a large influence on the estimated return to education.¹

The most noticeable result is probably the increase in the JIVE coefficient's standard error by factor 12 when controlling for age and age squared in a specification without further covariates, leading to a decrease of almost factor 10 in terms of the t -statistic. From specification (2) to (4) we observe an increase in the JIVE standard error by factor 8, implying a decrease in the t -statistic by factor 8.5. In contrast, the TSLS coefficient's standard error increases only by factor 1.8 and the t statistic decreases by factor 2.1 when adding age and age squared to specification (1). From specification (2) to (4) we observe an increase in the TSLS standard error by factor 1.8 and a decrease in t -statistic by factor 2.4. Thus, the TSLS coefficients are statistically significant on a 5% level in all specifications, while the JIVE estimates are highly insignificant when controlling for age and age squared. The use of bootstrapped standard errors also does not change this tendency. In contrast to the TSLS estimator, bootstrapped standard errors for the JIVE estimator are very volatile in specifications including age and age squared and appear to be highly dependent on the random bootstrap replicate samples chosen.

The problem with the split sample instrumental variables (SSIV) estimator is that the coefficients and standard errors very much depend on the random sample split. Thus, they are very volatile. The reported SSIV and USSIV values in Table 1 are mean coefficients and standard errors from 500 estimations. What can be observed is that significance of SSIV coefficients is similar to what we observe for the LIML and JIVE estimator. The USSIV coefficients being statistically significant on the 5% level are quite close to the corresponding LIML coefficients.² Bootstrapped SSIV standard errors are not reported here as they are similar to the usual ones in randomly drawn samples as long as the SSIV coefficient is significant with usual standard errors. That is to say that in specification (3) and (4) bootstrapped SSIV standard errors are much higher than in specifications without age and age squared. Given these problems with SSIV, we focus our attention in the subsequent analysis on TSLS, LIML, and JIVE.

¹As JIVE1 and JIVE2 estimators lead to very similar coefficients and standard errors in all specifications, we will concentrate on the JIVE1 estimator when interpreting our findings.

² That result does not depend on the size of the sample being used in the first-stage and second-stage regression. A table reporting SSIV results from 10% till 90% sample splits is available upon request from the authors.

5 Implications of Heteroskedasticity

An additional criterion on which to evaluate estimators is their performance in the presence of heteroskedasticity in the error term. While under homoskedasticity LIML may be preferred to TSLS on consistency grounds when the number of instruments increases proportionally to the sample size (Bekker (1994)), conventional standard errors will not be correct. Bekker (1994) provide valid standard errors, which can differ substantially from the conventional standard errors even with a small number of instruments. In a recent working paper, however, Hausman et al. (2008) point out that LIML can become inconsistent if heteroskedasticity is present. They also derive an exact condition for LIML to be consistent. Chao and Swanson (2004) argue that in the context of instrumental variables regression with many weak instruments, JIVE coefficients are consistent and asymptotically normal when errors are heteroskedastic of unknown form.

We apply heteroskedasticity tests after OLS, TSLS and LIML regressions in order to get an idea of whether a potential heteroskedasticity problem could influence the estimators' performance. Following the most commonly used testing procedure proposed by White (1980), one would regress the squared OLS, TSLS and LIML residuals each on levels, squares, and cross products of all instruments including all exogenous regressors, giving a test statistic of $nR^2 \sim \chi_j^2$, where j is the number of linearly independent regressors within these artificial regressions. The idea behind the Breusch and Pagan (1979), White (1980), and Pagan and Hall (1983) test is that the errors are conditionally heteroskedastic if any of the exogenous variables can predict the squared residuals.

Nevertheless, we only have 40 degrees of freedom with age which we already use up when including only levels of instruments. Therefore, in Table 2 we report results from regressing squared OLS, TSLS and LIML residuals each on levels of all excluded and included instruments.³ Our results suggest that heteroskedasticity might be a problem especially in specifications (2) and (4), which include 8 regional, and race, smsa, married dummies as further controls. The null hypothesis of homoskedasticity cannot be rejected for TSLS and LIML on a 5% level in specifications (1) and (3). Although we cannot be sure of the presence of this type of heteroskedasticity causing the LIML estimator to become inconsistent in the AK data, we find some evidence for the presence of a heteroskedasticity problem of unknown form especially in our preferred specification (4).

³ Test decisions do not change when regressing the squared residuals on levels, squares, and cross products of the instruments and dropping those regressors, whose coefficients show the lowest significance, afterwards.

6 The Problem of First-Stage Outliers

Our earlier results indicated that there is a large difference between the estimated return to education using TSLS and the estimates using JIVE in our preferred specification controlling for age and age squared in AK's data. The TSLS estimator and JIVE are asymptotically equivalent, however, and large differences in coefficients and estimated standard errors are therefore likely to be due to different first-stage predictions of the endogenous years of schooling variable. We can summarize the influence of leaving out observation i on the estimate of years of education by comparing $\widehat{\text{educ}}_{iIV}$ to $\widehat{\text{educ}}_{iJ}$. Since for each observation i there is a different vector of first-stage JIVE coefficients $\hat{\pi}_{iJ}$, we follow a method proposed by Cook (1977) to provide a combination of information from $\hat{\pi}_{iJ}$ in a scalar. Cook's Distance (henceforth CD) for observation i can be interpreted as the ordinary Euclidean distance between $\widehat{\text{educ}}_{iIV}$ and $\widehat{\text{educ}}_{iJ}$ and gives the squared distance from $\hat{\pi}_{iIV}$ to $\hat{\pi}_{iJ}$ relative to the fixed geometry of $Z'Z$. It is defined by

$$\text{CD}_i = \frac{(\hat{\pi}_{iIV} - \hat{\pi}_{iJ})' Z' Z (\hat{\pi}_{iIV} - \hat{\pi}_{iJ})}{k \hat{\sigma}_{iIV}^2} = \frac{(\widehat{\text{educ}}_{iIV} - \widehat{\text{educ}}_{iJ})^2}{k \hat{\sigma}_{iIV}^2} \quad (9)$$

where the second equality holds if Z contains only excluded instruments. Although CD_i does not follow an F distribution, Cook and Weisberg (1982) point out the convertibility of CD_i to that familiar scale. If CD_i equals the $1 - \alpha$ value of F_{n-k}^k , then deletion of the i^{th} observation would move $\hat{\pi}$ to the edge of a $1 - \alpha\%$ confidence ellipsoid relative to $\hat{\pi}$.

We use CD to identify outliers conditioning only on the excluded instruments and focus on estimation results from specifications (2) and (4). We calculate CD on the basis of a regression of education on only excluded instruments from the first-stage regression of each specification. In Figure 1 we plot the distribution of CD against years of schooling for the case with 28 excluded instruments (the graph for the case with 30 excluded instruments is quite similar); the line of observations with relatively small CD are the observations for which none of the dummy variables of the excluded instruments is equal to 1. From Figure 1 it is clear that CD is the larger the more extreme the value of years of education of person i is. In other words, the correlation between quarter of birth and years of schooling is especially weak for people with either a very low or a very high number of years of education. This is not surprising, because compulsory schooling laws are likely to be binding only on individuals who received somewhere around 11 years of schooling. For those groups of observations having a substantial influence on both the estimate of π and the fitted value $\widehat{\text{educ}}$, CD will be larger than for others. Deletion of those observations may lead to important changes in first-stage predictions and therefore in

estimated second-stage coefficients. Typically, in very large data sets the cases with the largest CDs will be of interest. To investigate the influence of those observations, we experiment with dropping different shares of the sample with the largest CDs and recompute TSLS and JIVE return to education coefficients to examine how those estimates change.

From Figure 2 it becomes clear that the distribution of years of education among the 2% sample of all observations with the largest values of CD looks quite different from the full sample distribution. Deleting observations with high CD values coincides with the deletion of extreme years of schooling observations, but not *vice versa*. In other words, for those observations the difference between $\hat{\pi}_{IV}$ and $\hat{\pi}_{iJ}$ is relatively large and quarter of birth performs particularly poor in explaining years of education.

Cook's Distance of observation i is determined by two components. The first is the standardized residual r_i reflecting the lack of fit of the model at the i^{th} case and the second is the location of Z_i relative to \bar{Z} over all observations, where Z in our case contains information on the quarter and year of birth. Weisberg (2005) point out that CD can numerically identically to equation (9) be calculated by

$$CD_i = \frac{1}{k} r_i^2 \frac{Z_i(Z'Z)^{-1}Z'_i}{1 - Z_i(Z'Z)^{-1}Z'_i} = \frac{1}{k} \frac{\hat{v}_i^2}{\hat{\sigma}_{IV}^2(1 - Z_i(Z'Z)^{-1}Z'_i)} \frac{Z_i(Z'Z)^{-1}Z'_i}{1 - Z_i(Z'Z)^{-1}Z'_i}, \quad (10)$$

where \hat{v}_{iIV} is the i^{th} residual from TSLS first-stage estimation. In Figure 3, we plot the standardized residual from the regression of years of education on the 28 year of birth \times quarter of birth instruments, sorted ascendingly by CD_i . There is a clear relationship. In Figure 4 we also plot the leverage of each observation, $Z_i(Z'Z)^{-1}Z'_i$, against CD_i . Here we find essentially no relationship, strongly suggesting that standardized residuals mainly determine CD in our case.

As first-stage JIVE predictions use a different vector $\hat{\pi}_{iJ}$ for each observation i and $\widehat{\text{educ}}_J$ depends on how π is estimated, large differences of $\hat{\pi}_{iJ}$ dependent on the left-out observation will lead to large estimated standard errors of JIVE coefficients. From Table 3 we can see that deletion of 2% of those observations with the largest CDs, we end up with highly significant and reasonable looking JIVE return to schooling coefficients. The JIVE return to schooling coefficients' standard errors decrease remarkably when those observations are deleted for which the instruments are especially poor, and the first-stage F statistics also indicate that any finite-sample bias is also substantially reduced.⁴ We hypothesize first-stage outliers will also

⁴We should note that we obtain very similar results if we also condition on the year-of-birth dummies and age and age squared. If we additionally include the other exogenous covariates in the regression we use to

be a serious problem for the SSIV estimator. As we have seen, the sample split is random and thus the coefficients depend very much on the randomly chosen sample.

In very large data sets it can easily happen that a group of cases is influential *en bloc* although this influence measured through CD is small when cases are examined individually. In our application, good candidates for exclusion out of the sample are those with the largest standardized residuals what coincides with a relatively large CD. While deletion of only one observation at a time in JIVE first-stage estimation does not seem to have substantial influence on the estimation of π measured by CD_i , the estimated variance of JIVE coefficients is substantially inflated by the extreme years of schooling observations as a group for which identification is extremely weak. Unfortunately, although our diagnostic analysis finds an outlier problem, it does not tell us what to do next. Sample truncation will probably affect other variables distributions and therefore influence the coefficient estimates. In terms of the covariates in specifications (2) and (4), observations with very low years of education are e.g. more likely to be black and not married. By dropping those observations from the sample we alter the distribution of those covariates. We interpret our findings as evidence for AIK-JIVE coefficients' standard errors to be especially high when identification is weak and a group of observations has substantial influence on the estimation of π and on the first-stage fitted values $\widehat{\text{educ}}$.

In order to shed further light on large estimated JIVE standard errors in specifications where identification is very weak, it is worth looking at the different components of the estimated coefficients' variance-covariance matrices. Let X be the matrix of structural equation regressors. Then Z is the matrix of instruments and \hat{X}_J is the matrix used in the second step of JIVE estimation with i^{th} row $Z_i \hat{\pi}(i) = Z_i (Z(i)' Z(i))^{-1} (Z(i)' X(i))$. The TSLS and LIML standard errors are estimated by taking the square root of the corresponding diagonal entry of

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2 [X' (I - \hat{\kappa} Z (Z' Z)^{-1} Z') X]^{-1} , \quad (11)$$

where $\hat{\sigma}^2$ and $\hat{\kappa}$ are calculated differently for the two estimators. The estimated TSLS coefficients' variance-covariance matrix results from setting $\hat{\kappa}_{IV} = 1$ and plugging in $\hat{\sigma}_{IV}^2 = \hat{u}'_{IV} \hat{u}_{IV} / N$ due to the large sample size. The residuals \hat{u}_{IV} are taken from the second-stage regression of TSLS estimation. In order to get the estimated variance-covariance matrix for the LIML coefficients, $\hat{\sigma}_L^2 = \hat{u}'_L \hat{u}_L / N$ is calculated using residuals from the LIML regression. To see how $\hat{\kappa}_L$ is estimated, we split X into a possibly endogenous part X_1 and an exogenous part X_2 . Fur-

identify the first-stage outliers, however, we obtain similar results to those in Table 1 for TSLS and JIVE. This is because race and marital status, in particular, also predict low and high levels of education.

thermore, we define $M_Z = I - Z(Z'Z)^{-1}Z'$ and $M_{X_2} = I - X_2(X_2'X_2)^{-1}X_2'$. In the application considered here, X_1 is years of education and X_2 contains all exogenous regressors. $\hat{\kappa}_L$ is the smallest eigenvalue of

$$[(y, X_1)'M_{X_2}(y, X_1)][(y, X_1)'M_Z(y, X_1)]^{-1}, \quad (12)$$

which can be characterized similarly by

$$\hat{\kappa}_L = \min_{\mu} \frac{\mu'(y, X_1)'M_{X_2}(y, X_1)\mu}{\mu'(y, X_1)'M_Z(y, X_1)\mu}. \quad (13)$$

The JIVE coefficients' standard errors are calculated from

$$\widehat{\text{Var}}(\hat{\beta}_J) = \hat{\sigma}_J^2 \left[X' \hat{X}_J (\hat{X}_J' \hat{X}_J)^{-1} \hat{X}_J' X \right]^{-1}, \quad (14)$$

where the residuals \hat{u}_J used to estimate $\hat{\sigma}_J^2 = \hat{u}_J' \hat{u}_J / N$ are taken from the JIVE regression.

Theil (1953) derives the difference between his TSLS estimator and the LIML estimator to be $o_p(1/\sqrt{n})$. In other words, LIML and TSLS have the same asymptotic distribution and $\sqrt{n}(\hat{\kappa} - 1) \rightarrow 0$ as $n \rightarrow \infty$ as noted by Anderson (2005) and Imbens and Wooldridge (2007).

As all estimated variance-covariance matrices are positive definite, we can apply theorem A.12 of Greene (2003) to the AK application to find $[\widehat{\text{Var}}(\hat{\beta}_J) - \widehat{\text{Var}}(\hat{\beta}_{IV})]$, $[\widehat{\text{Var}}(\hat{\beta}_J) - \widehat{\text{Var}}(\hat{\beta}_L)]$, and $[\widehat{\text{Var}}(\hat{\beta}_L) - \widehat{\text{Var}}(\hat{\beta}_{IV})]$ to be positive definite matrices.⁵ This results are independent of the specification used. Every eigenvalue of $\widehat{\text{Var}}(\hat{\beta}_J)$ is larger than the corresponding eigenvalue of $\widehat{\text{Var}}(\hat{\beta}_{IV})$ when both sets of eigenvalues are ordered from largest to smallest. The same holds true when comparing JIVE to LIML and LIML to TSLS.

For an invertible matrix the sum of its eigenvalues is equal to its trace. Therefore, for all considered estimators it holds that the sum of estimated variances of each component of $\hat{\beta}$ is equal to the sum of eigenvalues of the variance-covariance matrix of coefficients, i.e.

$$\sum_{j=1}^k \widehat{\text{Var}}(\hat{\beta}_j) = \text{tr}(\widehat{\text{Var}}(\hat{\beta})) = \sum_{l=1}^k \lambda_l, \quad (15)$$

where λ_l is the l^{th} eigenvalue of $\widehat{\text{Var}}(\hat{\beta})$. As can be seen from Table 4, the relation between traces of estimated coefficients' variance-covariance matrices is $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_J)) > \text{tr}(\widehat{\text{Var}}(\hat{\beta}_L)) >$

⁵ When using 28 instruments in specification (4), the difference of the smallest eigenvalues of $\widehat{\text{Var}}(\hat{\beta}_L)$ and $\widehat{\text{Var}}(\hat{\beta}_{IV})$ equals -2.474×10^{-13} . As this value is almost equal to zero, we can at least consider $\widehat{\text{Var}}(\hat{\beta}_L) - \widehat{\text{Var}}(\hat{\beta}_{IV})$ to be positive semidefinite.

$\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{IV}}))$ in all specifications. Furthermore, when using the full sample $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{J}}))$ is extremely high compared to $\text{tr}(\widehat{\text{Var}}(\hat{\beta}))$ of other estimators and to $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{J}}))$ in specifications (1) and (2).

Although the estimated error variance of the TSLS estimator, $\hat{\sigma}_{\text{IV}}^2$, is smaller than the estimated error variance of the LIML estimator, $\hat{\sigma}_{\text{L}}^2$, and $\hat{\sigma}_{\text{L}}^2$ is smaller than the estimated error variance of the JIVE estimator, $\hat{\sigma}_{\text{J}}^2$, in each specification, this alone will probably not be sufficient to explain the relation between the estimated variance-covariance matrices. Thus, it is worth to have a closer look on differences in the smallest eigenvalues of the matrices being defined by $A \equiv \hat{\sigma}^2[\widehat{\text{Var}}(\hat{\beta})]^{-1}$ for any one estimator. This is also the inverse of the second factor on the right hand side of equations (11) and (14). As we know about the equality described in equation (15) and that the eigenvalues of A are reciprocal to the eigenvalues of $A^{-1} = \widehat{\text{Var}}(\hat{\beta})/\hat{\sigma}^2$, a smallest eigenvalue of A close to zero will inflate $\sum_{j=1}^k \widehat{\text{Var}}(\hat{\beta}_j)$.

In the top panel of Table 4 we present the trace, error variance, and smallest root for the TSLS, LIML, and JIVE estimates in Table 1. In the AK data, the smallest eigenvalue of A_{IV} in specification (3) equals 2.24 and is 4.5 times smaller than in the specification with only year of birth dummies. $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{IV}}))$ increases by the same amount and is 4.5 times larger in specification (3) than in specification (1). The smallest characteristic root of A_{J} in specification (3) is around 0.03 and 186 times smaller than in specification (1), while $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{J}}))$ is 191.7 times larger. When including age and age squared in specification (2), the smallest eigenvalue of A_{IV} equals 2.09 and thus, decreases by factor 4.2 as $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{IV}}))$ increases by the same factor. In contrary, the smallest characteristic root of A_{J} equals 0.04 and is 88 times smaller in specification (4) than in specification (2). $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{J}}))$ increases by factor 88 when controlling for age and age squared in specification (2). When using the full sample, the smallest eigenvalues of A_{J} in the specifications including age and age squared are still positive, but quite close to zero. The characteristic roots of A_{IV} are all farther away from zero. This result coincides with $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{J}}))$ being very large while $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{IV}}))$ is comparatively small.

In the bottom panel of Table 4 we repeat the same exercise, recalculating the estimates using the truncated 98% sample from Table 3. Note that $\text{tr}(\widehat{\text{Var}}(\hat{\beta}))$ decreases for all estimators. A_{J} 's characteristic root is now much farther away from zero and $\text{tr}(\widehat{\text{Var}}(\hat{\beta}_{\text{J}}))$ is of reasonable size. Our eigenvalue analysis findings support an interpretation of JIVE coefficients' standard errors being inflated by outliers, for which the instruments are especially weak.

7 Conclusion

The TSLS estimator is known to be potentially seriously biased if identification is weak. Using well-known data from AK, we estimate the returns to education using a variety of estimators that might be preferred on prior grounds (LIML, JIVE, and SSIV). In our preferred specification that includes year-of-birth dummies, age, age squared, and further covariates, we find evidence highly variable results across the different estimation methods. In particular, results from SSIV and JIVE, which is the preferred estimator on *a priori* grounds with weak identification, are never statistically significant. We reject the LIML and TSLS estimates as likely to be inconsistent because of heteroskedasticity.

We obtain plausible and statistically significant results from JIVE only from a truncated sample by dropping the most influential observations with either extremely few or many years of education; it is precisely these observations where the natural experiment that AK utilize for identification would be expected to be roughly binding. We conclude from this exercise that in the full sample JIVE coefficients' standard errors are inflated by those observations for which identification is especially weak. As sample truncation alters the distribution of other variables we cannot interpret return to education estimates from those estimations as generally applicable.

References

- Anderson, T. (2005). Origins of the Limited Information Maximum Likelihood and Two-Stage Least Squares Estimators. *Journal of Econometrics*, 127(1):1–16.
- Angrist, J., Imbens, G., and Krueger, A. (1999). Jackknife Instrumental Variables Estimation. *Journal of Applied Econometrics*, 14:57–67.
- Angrist, J. and Krueger, A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics*, 106(4):979–1014.
- Angrist, J. and Krueger, A. (1995). Split-Sample Instrumental Variables Estimates of the Return to Schooling. *Journal of Business and Economic Statistics*, 13(2):225–235.
- Bekker, P. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, 62(3):657–681.
- Bound, J., Jaeger, D., and Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.
- Breusch, T. and Pagan, A. (1979). A Simple Test for the Heteroskedasticity and Random Coefficient Variation. *Econometrica*, 47:1287–1297.
- Card, D. (1999). The Causal Effect of Education on Earnings. *Handbooks in Economics*, 5(3A):1801–1864.
- Chao, J. and Swanson, N. (2004). Estimation and Testing Using Jackknife IV in Heteroskedastic Regressions with Many Weak Instruments. *Rutgers University Economics Working Paper No. 2004-20*.
- Chao, J. and Swanson, N. (2005). Consistent Estimation with a Large Number of Weak Instruments. *Econometrica*, 73(5):1673–1692.
- Chao, J., Swanson, N., Hausman, J., Newey, W., and Woutersen, T. (2007). Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments. Technical report, Working Paper, Rutgers.
- Cook, R. (1977). Detection of Influential Observation in Linear Regression. *Techometrics*, 19:15–18.

- Cook, R. and Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall New York.
- Greene, W. (2003). *Econometric Analysis* (5th).
- Hausman, J., Newey, W., Woutersen, T., Chao, J., and Swanson, N. (2008). Instrumental Variable Estimation with Heteroskedasticity and Many Instruments. *Working Paper*.
- Imbens, G. and Wooldridge, J. (2007). Weak Instruments and Many Instruments. *Lecture Notes 13, NBER Summer 2007*.
- Mikusheva, A. (2007). Uniform Inferences in Econometrics. *Chapter 3, PhD Thesis, Harvard University, Department of Economics*.
- Pagan, A. and Hall, A. (1983). Diagnostic Tests as Residual Analysis. *Econometric Reviews*, 2(2):159–218.
- Staiger, D. and Stock, J. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586.
- Stock, J., Wright, J., and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics*, 20(4):518–529.
- Stock, J. and Yogo, M. (2002). Testing for Weak Instruments in Linear IV Regression. *NBER Working Paper*.
- Theil, H. (1953). Repeated Least Squares Applied to Complete Equation Systems. *The Hague: Central Planning Bureau*.
- van der Klaauw, W. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach. *International Economic Review*, 43(4):1249–1287.
- Weisberg, S. (2005). *Applied Linear Regression*. John Wiley & Sons New York.
- White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4):817–838.

Table 1

Estimation of the Return to Education for Men Born 1930-1939

Years of education	(1)	(2)	(3)	(4)
OLS Coefficient	0.0711	0.0632	0.0711	0.0632
Std. Error	0.0003	0.0003	0.0003	0.0003
Bootstrapped Std. Error ^(a)	0.0004	0.0004	0.0004	0.0004
TOLS Coefficient	0.0891	0.0806	0.0760	0.0600
Std. Error	0.0161	0.0164	0.0290	0.0290
Bootstrapped Std. Error ^(a)	0.0168	0.0154	0.0282	0.0266
First-stage F	4.91	4.75	1.61	1.61
Partial R ² *100	0.04	0.04	0.01	0.01
LIML Coefficient	0.0929	0.0838	0.0810	0.0574
Std. Error	0.0177	0.0179	0.0412	0.0388
Bootstrapped Std. Error ^(a)	0.0248	0.0226	0.6243	0.2116
SSIV Coefficient ^(b)	0.0638	0.0570	0.0217	0.0144
Std. Error ^(b)	0.0223	0.0223	0.0350	0.0346
USSIV Coefficient ^(b)	0.0941	0.0853	0.0791	0.0543
Std. Error ^(b)	0.0317	0.0326	0.4416	0.4557
JIVE1 Coefficient	0.0959	0.0904	0.1150	0.0860
Std. Error	0.0222	0.0258	0.2644	0.2110
Bootstrapped Std. Error ^(a)	0.0213	0.0242	0.5337	2.1461
JIVE2 Coefficient	0.0959	0.0904	0.1155	0.0862
Std. Error	0.0222	0.0258	0.2643	0.2094
Bootstrapped Std. Error ^(a)	0.0214	0.0244	0.2298	2.7670
Age, age squared	No	No	Yes	Yes
Year of birth dummies	Yes	Yes	Yes	Yes
Other covariates ^(c)	No	Yes	No	Yes
Number of instruments	30	30	28	28

Source: Authors' tabulations from 1980 U.S.Census used by Angrist and Krueger (1991).

Note: Sample size is 329,509. The 30 instruments are a set of 3 quarter-of-birth times 10 year-of-birth interactions. The dependent variable is the log of weekly earnings. Age and age squared are measured in quarters of years.

a: 500 bootstrap replications.

b: Mean from 500 estimations.

c: 8 regional, and race, smsa, married dummies as further controls.

Table 2

Heteroskedasticity Tests for the Estimation of the Return to Education

Years of education	(1)	(2)	(3)	(4)
OLS Coefficient	0.0711	0.0632	0.0711	0.0632
Std. Error	0.0003	0.0003	0.0003	0.0003
White nR^2	150.52	744.40	153.80	747.62
(degrees of freedom, p-value)	(10, 0.00)	(21, 0.00)	(12, 0.00)	(23, 0.00)
TSLS Coefficient	0.0891	0.0806	0.0760	0.0600
Std. Error	0.0161	0.0164	0.0290	0.0290
White nR^2	47.98	750.61	45.79	737.23
(degrees of freedom, p-value)	(39, 0.15)	(50, 0.00)	(39, 0.21)	(50, 0.00)
First-stage White nR^2	647.36	3273.70	647.36	3273.70
(degrees of freedom, p-value)	(39, 0.00)	(50, 0.00)	(39, 0.00)	(50, 0.00)
LIML Coefficient	0.0929	0.0838	0.0810	0.0574
Std. Error	0.0177	0.0179	0.0412	0.0388
White nR^2	48.78	753.97	46.48	735.58
(degrees of freedom, p-value)	(39, 0.14)	(50, 0.00)	(39, 0.19)	(50, 0.00)
Age, age squared	No	No	Yes	Yes
Year of birth dummies	Yes	Yes	Yes	Yes
Other covariates ^(a)	No	Yes	No	Yes
Number of instruments	30	30	28	28

Source: Authors' tabulations from 1980 U.S.Census used by Angrist and Krueger (1991).

Note: Sample size is 329,509. The 30 instruments are a set of 3 quarter-of-birth times 10 year-of-birth interactions. The dependent variable is the log of weekly earnings. Age and age squared are measured in quarters of years.

a: 8 regional, and race, smsa, married dummies as further controls.

Figure 1

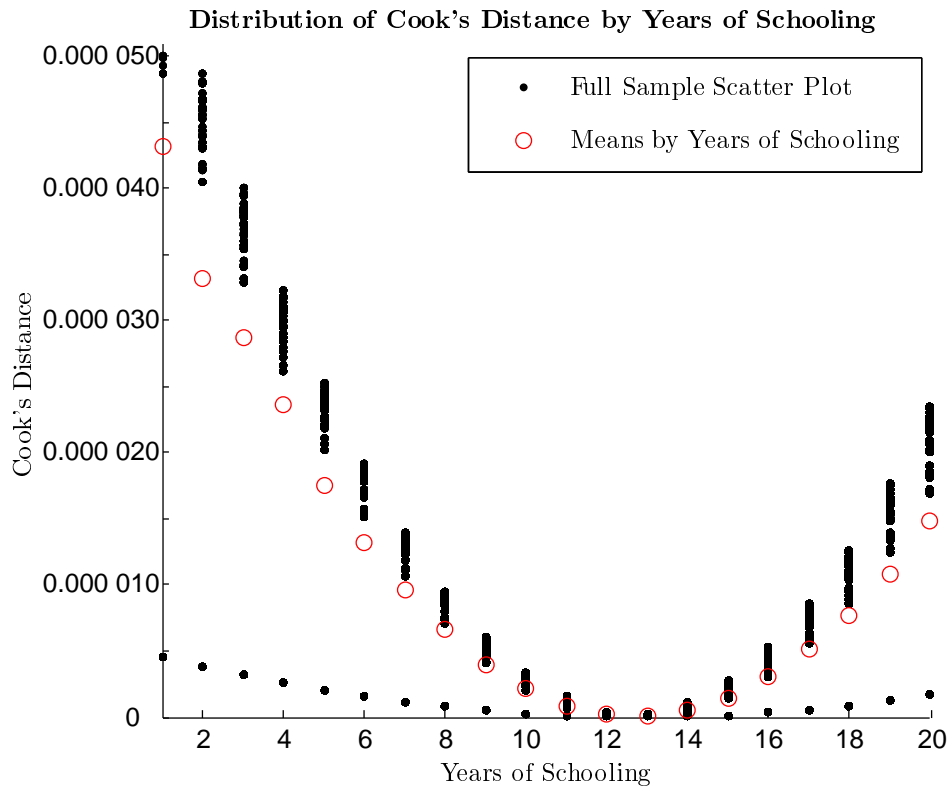


Figure 2

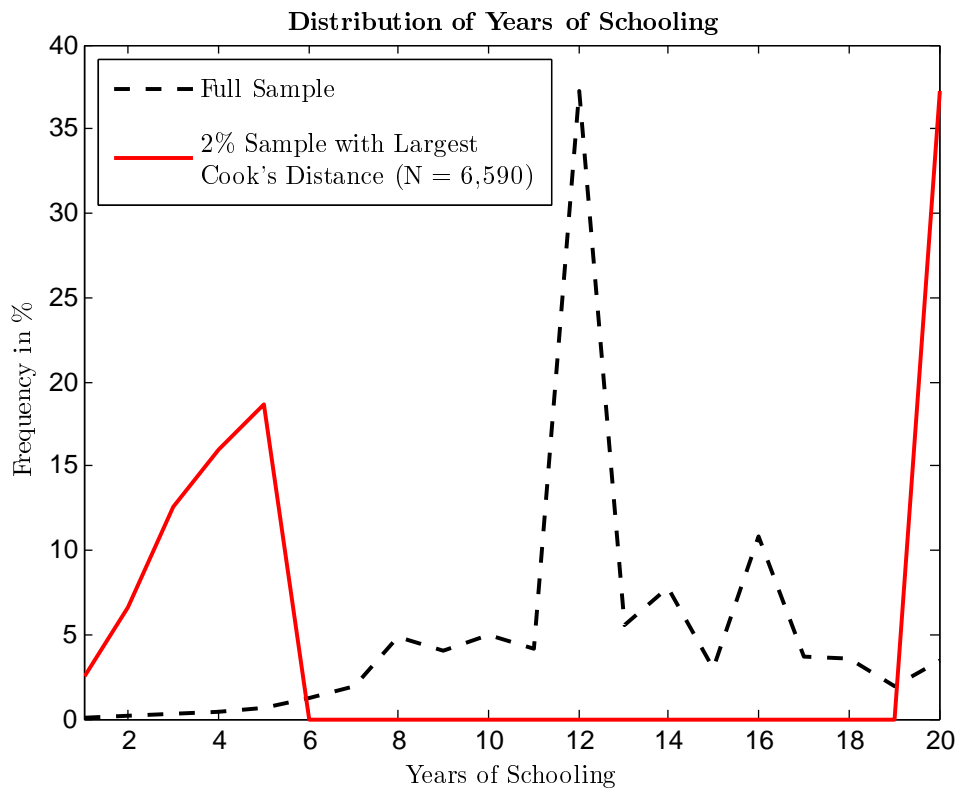


Figure 3

Distribution of Standardized Residuals sorted Ascendingly by Cook's Distance

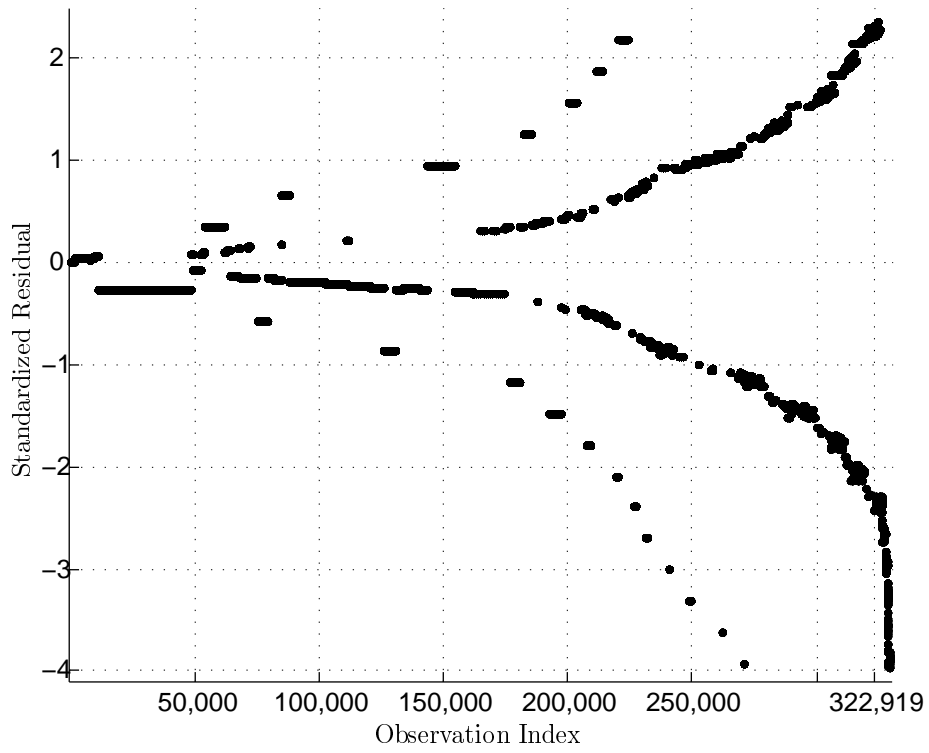


Figure 4

Distribution of Leverage sorted Ascendingly by Cook's Distance

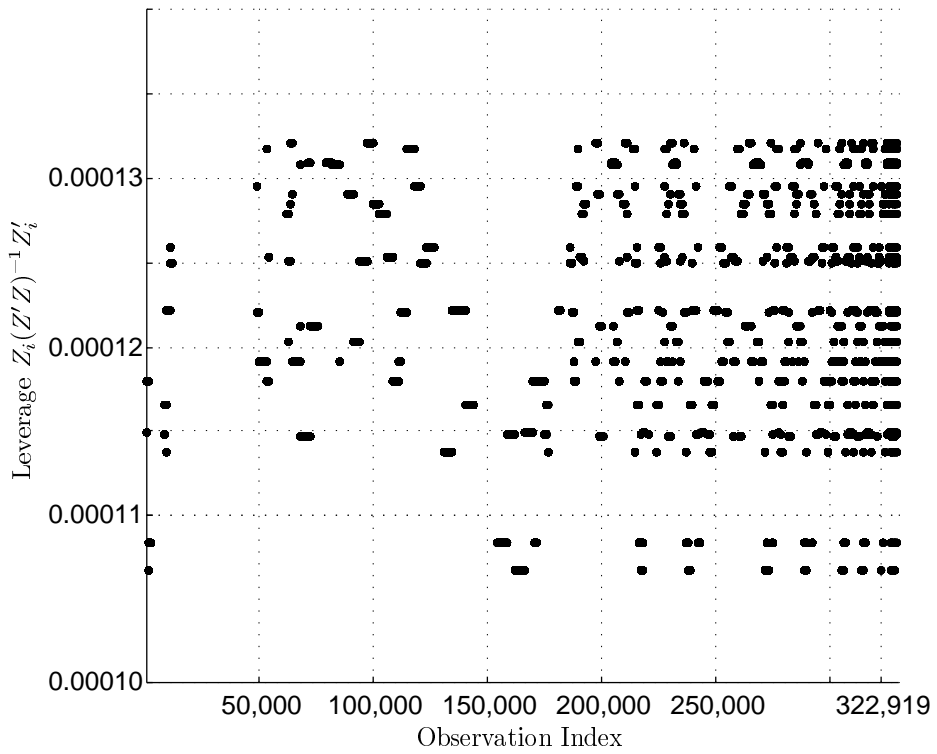


Table 3

Estimation of the Return to Education for Men Born 1930-1939

Years of education	(2)		(4)	
	TSLS	JIVE1	TSLS	JIVE1
Full sample (N = 329,509)				
Coefficient	0.0806	0.0904	0.0600	0.0860
Std. Error	0.0164	0.0258	0.0290	0.2110
First stage:				
F statistic ^(b)	4.7474	4.7474	1.5871	1.6131
99% sample (N = 326,213) ^(a)				
Coefficient	0.0761	0.0886	0.0616	0.0597
Std. Error	0.0201	0.0408	0.0193	0.0366
First stage:				
F statistic ^(b)	3.3785	3.3785	3.8839	3.8841
98% sample (N = 322,919) ^(a)				
Coefficient	0.0747	0.0765	0.0735	0.0761
Std. Error	0.0111	0.0130	0.0137	0.0177
First stage:				
F statistic ^(b)	11.6307	11.6307	8.1116	8.1127
Age, age squared	Yes	Yes	Yes	Yes
Year of birth dummies	Yes	Yes	Yes	Yes
Other covariates ^(c)	No	No	Yes	Yes
Number of instruments	30	30	28	28

Source: Authors' tabulations from 1980 U.S.Census used by Angrist and Krueger (1991).

Note: The 30 instruments are a set of 3 quarter-of-birth times 10 year-of-birth interactions. The dependent variable is the log of weekly earnings. Age and age squared are measured in quarters of years. Cook's Distance is calculated from regressing years of schooling on the excluded instruments of the corresponding TSLS first stage regression.

a: Sample is truncated by dropping observations with largest Cook's Distance.

b: F statistic of excluded instruments.

c: 8 regional, and race, smsa, married dummies as controls.

Table 4

Properties of the Estimated Variance-Covariance Matrices

	(1)	(2)	(3)	(4)
Full Sample (N = 329,509)				
TOLS Estimation				
Trace of estimated $\text{Var}(\hat{\beta})$	0.0411	0.0449	0.1868	0.1896
Estimated Error Variance	0.4100	0.3878	0.4068	0.3849
Smallest Root ^(a)	10.0371	8.7268	2.2359	2.0850
LIML Estimation				
Trace of estimated $\text{Var}(\hat{\beta})$	0.0498	0.0533	0.3702	0.3332
Estimated Error Variance	0.4116	0.3890	0.4076	0.3851
Smallest Root ^(a)	8.3060	7.3521	1.1157	1.1733
JIVE1 Estimation				
Trace of estimated $\text{Var}(\hat{\beta})$	0.0780	0.1105	14.9497	9.6703
Estimated Error Variance	0.4131	0.3953	0.4273	0.3900
Smallest Root ^(a)	5.3133	3.5926	0.0286	0.0404
98% sample (N = 322,919) ^(b)				
TOLS Estimation				
Trace of estimated $\text{Var}(\hat{\beta})$	0.0195	0.0208	0.0426	0.0452
Estimated Error Variance	0.4020	0.3808	0.4033	0.3808
Smallest Root ^(a)	20.8527	18.7371	10.6813	9.4527
LIML Estimation				
Trace of estimated $\text{Var}(\hat{\beta})$	0.0209	0.0221	0.0459	0.0485
Estimated Error Variance	0.4021	0.3810	0.4036	0.3809
Smallest Root ^(a)	19.4611	17.5745	9.8462	8.7470
JIVE1 Estimation				
Trace of estimated $\text{Var}(\hat{\beta})$	0.0247	0.0283	0.0582	0.0705
Estimated Error Variance	0.4162	0.3991	0.4184	0.3943
Smallest Root ^(a)	16.9784	14.3011	7.8508	6.0226
Age, age squared	No	No	Yes	Yes
Year of birth dummies	Yes	Yes	Yes	Yes
Other covariates ^(c)	No	Yes	No	Yes
Number of instruments	30	30	28	28

Source: Authors' tabulations from 1980 U.S.Census used by Angrist and Krueger (1991).

Note: The 30 instruments are a set of 3 quarter-of-birth times 10 year-of-birth interactions. The dependent variable is the log of weekly earnings. Age and age squared are measured in quarters of years. Cook's Distance is calculated from regressing years of schooling on the excluded instruments of the corresponding TOLS first stage regression.

a: Smallest eigenvalue of the inverse of the estimated variance-covariance matrix times the estimated error variance.

b: Sample is truncated by dropping observations with largest Cook's Distance.

c: 8 regional, and race, smsa, married dummies as controls.