I benefited from discussions with Thomas Bearpark and Eric Qian. All errors are my own.

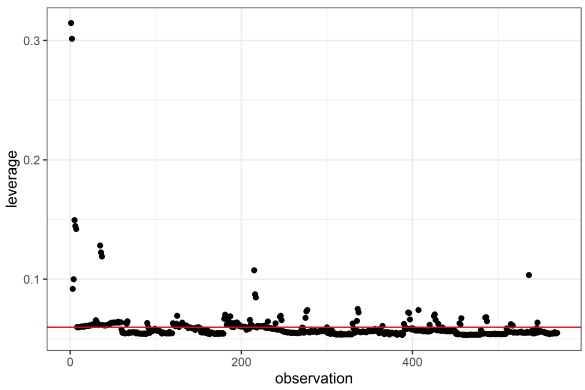
Question 1

Following Meng, Qian and Yared (2015) we estimate the regression of

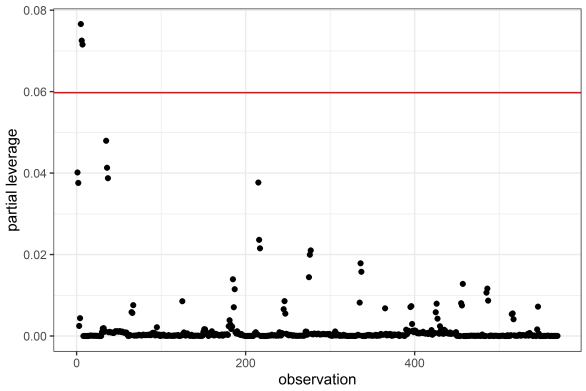
$$y_{it} = X_{it} \times D_{it}\beta_1 + X_{it} \times (1 - D_{it})\,\beta_2 + \mathbf{W}_{it}'\boldsymbol{\gamma} + \alpha_t + \varepsilon_{it}, \tag{1}$$

where y_{it} is the log-mortality rate on log-grain production in famine and non-famine periods, year fixed effects, and a vector of covariates containing log of total and urban population. In the regression above β_1 and β_2 can be interpreted as the elasticities of the mortality rate to grain production in famine and non-famine periods, respectively. Table 1 reports the point estimates together with confidence intervals computed using various techniques.
Overall, clustering increases the length of the estimated confidence intervals, whereas correcting for the presence of outliers does not affect the results. Indeed, leverage and partial leverage are low for almost all units across various specifications (Figure 3). This does not change our conclusion regarding the effect of food production on mortality in non-famine years, as we always fail to reject the null of a non-zero effect. On the contrary, when it comes to assess whether food production had a statistically significant effect on mortality in famine years our answer is less clear cut. Controlling for correlated shocks at the province level makes the estimate of β_1 not statistically significant with 2 of the 5 different standard errors we use (full sample) and with all procedures (restricted sample). It also seems that clustering standard errors is the right thing to do, as provinces are likely to be serially correlated and we only observe a subset of the superpopulation of Chinese provinces.

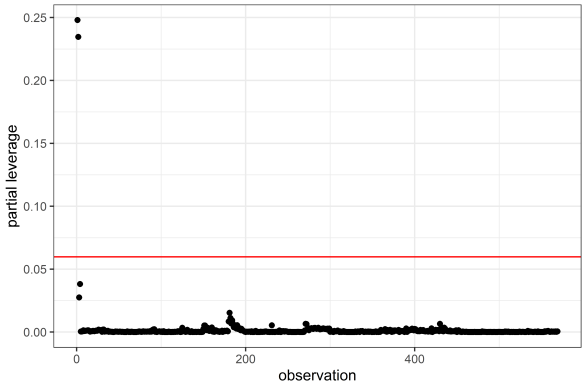
Figure 1: *Various measures of leverage (full sample).*



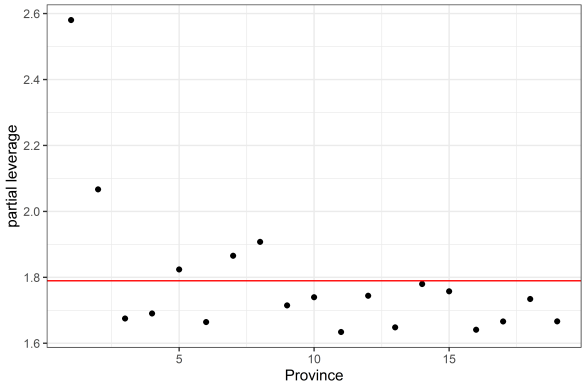
(i) *Unit leverage* H.



(ii) *Partial leverage* $\widetilde{H}_{X\times D}$.



(iii) *Partial leverage* $\widetilde{H}_{X\times(1-D)}$.



(iv) *Partial Cluster-leverage* $\widetilde{H}_{g,X\times D}$.

Notes: Leverage is computed as $H = X\left(X'X\right)^{-1}X'$, where X is the design matrix used in the regression . The partial leverage of the variable Z is defined as $\widetilde{H}_Z = \widetilde{Z}\left(\widetilde{Z}'\widetilde{Z}\right)^{-1}\widetilde{Z}'$, where $\widetilde{Z}$ is the residual of the projection of Z on all the other columns of X. We follow MacKinnon et al. (2022) and compute the cluster leverage as the nuclear norm of H_g. The horizontal solid red line indicates the ideal benchmark of balanced leverage.

Table 1: *Point estimates of β_1 and β_2 with 95% confidence intervals.*

***Sample**:* 1953-1982		*Food Production famine*		*Food Production non-famine*	
Point Estimate:		0.141		-0.007	
95% Confidence Interval:		*Lower Bound*		*Upper Bound*	
Heteroskedasticity-Robust					
	HC1	0.057	0.225	-0.032	0.019
	HC2	0.056	0.227	-0.034	0.021
	Satt	0.052	0.231	-0.040	0.026
	bootstrap (np)	0.051	0.232	-0.039	0.025
	bootstrap (pct)	0.062	0.241	-0.031	0.024
Cluster-Robust					
	CR1	0.024	0.258	-0.059	0.046
	CR2	0.023	0.260	-0.066	0.052
	Satt	-0.006	0.288	-0.113	0.099
	bootstrap (np)	-0.004	0.286	-0.100	0.086
	bootstrap (pct)	0.026	0.231	-0.056	0.055
***Sample**:* 1953-1965					
Point Estimate:		0.098		-0.005	
95% Confidence Interval:		*Lower Bound*		*Upper Bound*	
Heteroskedasticity-Robust					
	HC1	0.013	0.183	-0.029	0.019
	HC2	0.011	0.184	-0.030	0.019
	Satt	0.008	0.187	-0.038	0.027
	bootstrap (np)	0.006	0.189	-0.037	0.027
	bootstrap (pct)	0.015	0.192	-0.028	0.021
Cluster-Robust					
	CR1	-0.007	0.203	-0.044	0.033
	CR2	-0.011	0.207	-0.052	0.041
	Satt	-0.034	0.229	-0.102	0.092
	bootstrap (np)	-0.047	0.242	-0.102	0.091
	bootstrap (pct)	-0.013	0.177	-0.042	0.038

Notes: HC1 uses the Eicker-Huber-White (EHW) standard errors, *HC2* uses the EHW standard errors with the MacKinnon and White (1985) correction for high-leverage observations; *Satt* uses the HC2 standard errors with the Satterthwaite (1946) degrees of freedom approximation; *bootstrap (np)* uses the standard errors obtained estimating $\widehat{\beta}_j, j = 1, 2$ in 50.000 draws of a non-parametric bootstrap; *bootstrap (pct)* constructs confidence interval by computing the robust-t statistic (with HC1 standard errors) in 50.000 draws of a percentile bootstrap; *CR1* uses the Liang and Zeger (1986) cluster-robust standard errors; *CR2* uses the Bell and McCaffrey (2002) cluster-robust standard errors with the correction for high-leverage clusters. The last two rows rely on a block bootstrap where provinces have been sampled instead of sampling the single unit of observation.

Question 2 - Critical Reading, Kean and Neal (2021)

Summary. The paper discusses about the finite sample properties of the two-stage least squares estimator (TSLS) and addresses the issue of inference in such context.

First, the authors survey the most popular results and suggestions in the literature on weak instrumental variables and then shows how conventional t-tests lead to misleading inference even when instruments are deemed as "strong" according to previous research. Indeed, correct nominal size of 5% is restored at values higher than 100 (shown in Lee, McCrary, Moreira and Porter 2021). This is mainly attributed to the well known fact that the finite sample distribution of TSLS is asymmetric and with fat tails.

Second, the authors focus on the fact that the asymmetry in the distribution has important consequences when it comes to address different null hypotheses (e.g. $H_0 : \beta < 0$ or $H_0 : \beta > 0$). Towards this goal, it is shown that the TSLS estimator has artificially low standard errors precisely when its point estimates are biased towards OLS. This implies that, depending on the direction of the bias, conventional t-tests are extremely low-powered against either positive or negative alternatives.

Third, it is shown that one-tailed tests have much greater size distortion than two-tailed tests. This happens because of the asymmetry of the TSLS distribution. To restore the traditional symmetry between two-tailed and one-tailed tests, independently of the degree of endogeneity, the population F-statistic must be in the thousands. The classical Anderson-Rubin (AR) test achieves such balance for a vastly smaller F.

Finally, the conditional t-test approach is shown to have both correct overall 5% rejection rate (same as AR) and maintain this property also when it comes to one-sided hypothesis tests. Indeed, such approach adjusts the critical values to take the non-normality and asymmetry of the TSLS distribution into account. However, since the AR is the UMP test in the class of two-tailed tests with symmetric critical values, there must be a trade-off. In particular, the authors suggest to use the AR when the true effect goes in the opposite direction to the OLS bias, and the conditional approach in the other case.

"First-stage F in the thousands". The critique that the authors make is relevant only if the degree of endogeneity is extremely high and a researcher is interested in a one-sided hypothesis test. In a sense, it embraces the attitude of creating prescriptions to practitioners that are valid even in the *worst case scenario* (Staiger and Stock 1997, Stock and Yogo 2005, Lee, McCrary, Moreira and Porter 2021). However, despite being an interesting theoretical point, this critique has little practical relevance in economic applications or whenever a researcher has knowledge about the magnitude/sign of β. Furthermore, in the just-identified IV case there exists a one-to-one mapping between β and the degree of endogeneity ρ. Hence, by restricting the value of β on a specific interval, it is also possible to bound the amount of possible endogeneity. Taking these facts into consideration sensibly reduces the range of applications to which the critique can be applied.

Takeaway. Robust estimators and testing procedures are usually appealing because they save the researcher from justifying why an empirical question has been answered in a certain manner. However, robustness comes at the expenses of efficiency. Hence, if a researcher possesses relevant information about the problem at hand, this should be used to alleviate such trade-off. In the last twenty years, the weak-IV literature has taken a conservative approach and suggested testing procedures that controlled size distortion even under the worst-case scenario. Even though these suggestions have wide applicability, they should not be naively implemented when additional structure can be imposed on the problem at hand, making the worst-case scenario less worse. In

this spirit, Angrist and Kolesár (2022) show that in the economic applications that they consider, $|\rho| < 0.5$ (Table 1). They also show that when $|| < 0.565$ the nominal 5% t-test under-rejects for any population value of F (see also Figure 2). At least in the applications considered in Angrist and Kolesár (2022), the advice given in Kean and Neal (2021) has a little role to play.

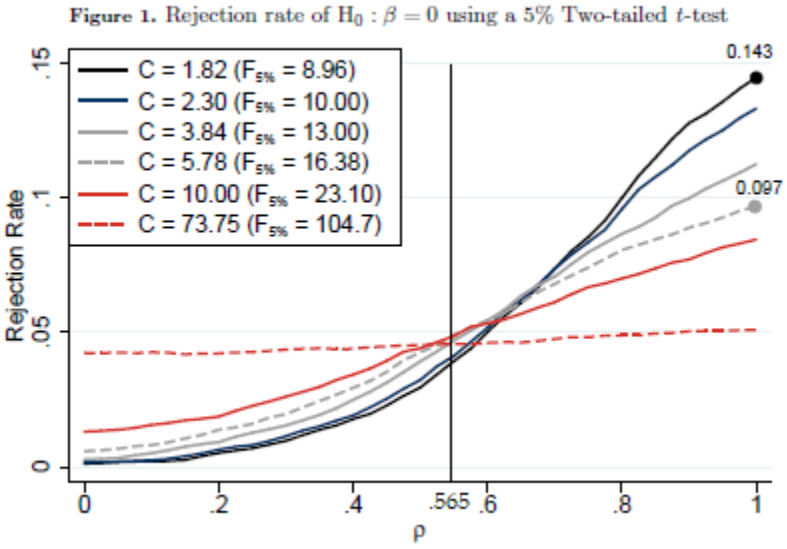


Figure 2: *Source: Kean and Neal (2021).*

Question 3

Table 2: *Estimated returns to schooling in different sub-samples.*

	Pacific	*mid-Atlantic*
Point Estimate:		
	0.13	-1.52
Confidence Interval:		
Wald	$[-0.02; 0.27]$	$[-20.67; 17.62]$
AR	$[-0.02; 0.38]$	$[-\infty; \infty]$
tF	$[-0.14; 0.39]$	$[-\infty; \infty]$
First-stage F	9.29	0.03

Table 2 reports the point estimate for returns of schooling in different sub-samples of the original Angrist and Krueger (1991) dataset. We can see that in the Pacific sub-sample, the first-stage F is around 9, suggesting that the traditional Wald confidence interval might be undercovering. Both the Anderson-Rubin and the tF procedure yield larger confidence interval. In the mid-Atlantic case, the instrument we are using is extremely weak and has a first-stage F that is almost 0. This induces the IV estimator to be extremely noisy (returns to schooling are indeed very negative) and the estimand of interest to be ultimately unidentifiable. Conventional Wald confidence interval ignore that in a region of the parameter space the model is weakly identified, thus they continue to be a bounded interval. However, the possibility of obtaining an infinite confidence set is a necessary condition for having a procedure robust to weak instruments (Dufour 1997). If instruments are weak, then the data contain little information about the coefficient of interest, resulting in infinite confidence sets. Both the Anderson-Rubin and the tF procedure have correct nominal coverage independently of the first-stage F, thus their expected length is infinite. In particular, this happens because a necessary and sufficient condition for these intervals to be bounded is the population F being such that $F < q_{1-\alpha}$.

Finally, if are willing to a priori restrict returns to education to lie between 0 and 0.25, we can bound the extent of the endogeneity we might face. Indeed, as shown in Angrist and Kolesár (2022), under homoskedasticity we can express the degree of endogeneity ρ as

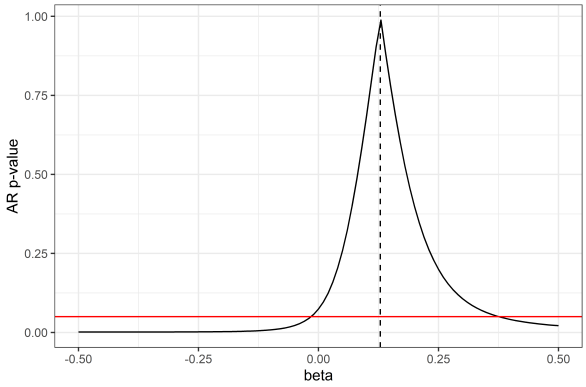
$$\rho \approx \frac{\sigma_D}{\sigma_Y}\left(\frac{\beta_{OLS} - \beta}{1 - R_p^2}\right),$$

where σ_D is the standard deviation of the education variable, σ_Y is the standard deviation of log-wages, and R_p^2 is the first-stage partial R-squared. This latter term is approximately 0 in both sub-samples, thus we can safely ignore it. Estimating the quantity above using sample analogs and plugging in 0 and 0.25 in place of β we get

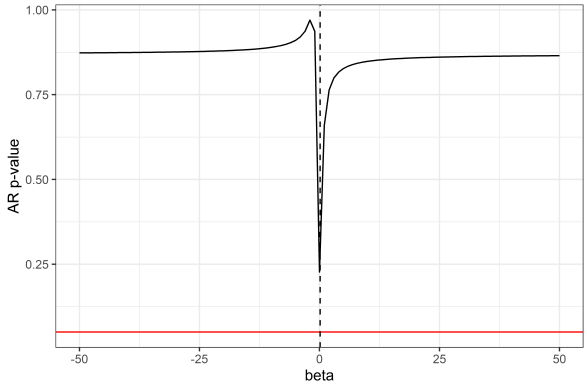
$$\widehat{\rho}_{\texttt{pacific}} \in [-0.889, 0.264]\,, \qquad \widehat{\rho}_{\texttt{mid-atlantic}} \in [-0.832, 0.361]\,.$$

According to these estimates and to Figure 1 in Angrist and Kolesár (2022), the traditional Wald confidence interval has maximum possible size distortion of 15%, calculated on the worst-case scenario of $E[F] = 1$.

Figure 3: *Anderson-Rubin confidence intervals in different sub-samples.*



(i) *Pacific division sub-sample.*



(ii) *mid-Atlantic division sub-sample.*

References

Angrist, J. D. and Krueger, A. B. (1991), 'Does compulsory school attendance affect schooling and earnings?', *The Quarterly Journal of Economics* **106**(4), 979–1014.

Angrist, J. and Kolesár, M. (2022), 'One instrument to rule them all: The bias and coverage of just-id iv'.

Bell, R. M. and McCaffrey, D. F. (2002), 'Bias reduction in standard errors for linear regression with multi-stage samples', *Survey Methodology* **28**(2), 169–182.

Dufour, J.-M. (1997), 'Some impossibility theorems in econometrics with applications to structural and dynamic models', *Econometrica: Journal of the Econometric Society* pp. 1365–1387.

Kean, M. and Neal, T. (2021), A practical guide to weak instruments, Working paper, SSRN.
URL: https://doi.org/10.2139/ssrn.3846841

Lee, D. S., McCrary, J., Moreira, M. J. and Porter, J. R. (2021), Valid t-ratio inference for iv, Technical report, National Bureau of Economic Research.

Liang, K.-Y. and Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

MacKinnon, J. G., Nielsen, M. Ã., Webb, M. D. et al. (2022), 'Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust'.

MacKinnon, J. G. and White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of econometrics* **29**(3), 305–325.

Meng, X., Qian, N. and Yared, P. (2015), 'The institutional causes of china's great famine, 1959–1961', *The Review of Economic Studies* **82**(4), 1568–1611.

Satterthwaite, F. E. (1946), 'An approximate distribution of estimates of variance components', *Biometrics bulletin* **2**(6), 110–114.

Staiger, D. and Stock, J. (1997), 'Instrumental variables regression with weak instruments', *Econometrica* **65**, 557–586.

Stock, J. and Yogo, M. (2005), *Asymptotic distributions of instrumental variables statistics with many instruments*, Vol. 6, Chapter.