

A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs

Seojeong (Jay) Lee^{*†}

September 28, 2015

Abstract

Under treatment effect heterogeneity, an instrument identifies the instrument-specific local average treatment effect (LATE). With multiple instruments, two-stage least squares (2SLS) estimand is a weighted average of different LATEs. In practice, a rejection of the over-identifying restrictions test can indicate that there are more than one LATE. What is often overlooked in the literature is that the postulated moment condition evaluated at the 2SLS estimand does not hold unless those LATEs are the same. If so, the conventional heteroskedasticity-robust variance estimator would be inconsistent, and 2SLS standard errors based on such estimators would be incorrect. I derive the correct asymptotic distribution, and propose a consistent asymptotic variance estimator by using the result of Hall and Inoue (2003, *Journal of Econometrics*) on misspecified moment condition models. This can be used to correctly calculate the standard errors regardless of whether there are more than one LATE or not.

Keywords: local average treatment effect, treatment heterogeneity, two-stage least squares, variance estimator, model misspecification

JEL Classification: C13, C31, C36

^{*}School of Economics, UNSW Business School, University of New South Wales, Sydney NSW 2052 Australia, email: jay.lee@unsw.edu.au, homepage: <https://sites.google.com/site/misspecified/>

[†]Conversations with Tue Gorgens motivated this paper. I also thank Bruce Hansen, Jack Porter, Han Hong, Toru Kitagawa, Sukjin Han, Ju Hyun Kim, Gigi Foster, Valentyn Panchenko, Denzil Fiebig, and Taehoon Kim as well as session participants at SETA 2014, KAEA-KEA, the Econometric Society World Congress 2015 conferences for their helpful discussions and suggestions.

1 Introduction

Since the series of seminal papers by Imbens and Angrist (1994), Angrist and Imbens (1995), and Angrist, Imbens, and Rubin (1996), the local average treatment effect (LATE) has played an important role in providing useful guidance to many policy questions. The key underlying assumption is treatment effect heterogeneity, i.e. each individual has a different causal effect of treatment on outcome. Assume a binary treatment, D_i , and an outcome variable Y_i . Let Y_{1i} and Y_{0i} denote the potential outcomes of individual i with and without the treatment, respectively. The heterogeneous individual treatment effect is $Y_{1i} - Y_{0i}$, but this cannot be identified because Y_{1i} and Y_{0i} are never observed at the same time. Instead, the average treatment effect (ATE), $E[Y_{1i} - Y_{0i}]$, may be policy-relevant. However, unless the treatment status is randomly assigned, a naive estimate of ATE is likely to be biased because of selection into treatment.

Instrumental variables are used to overcome this endogeneity problem. If an instrument Z_i which is independent of Y_{1i} and Y_{0i} , and correlated with the treatment D_i is available, then ATE of those whose treatment status can be changed by the instrument, thus the *local* ATE, is identified. Assume Z_i is binary and define D_{1i} and D_{0i} be i 's treatment status when $Z_i = 1$ and $Z_i = 0$, respectively. The LATE theorem of Imbens and Angrist (1994) shows that

$$\frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}]. \quad (1.1)$$

That is, the instrumental variables (IV) estimand (or the Wald estimand) is equal to the ATE for a subpopulation such that $D_{1i} > D_{0i}$, which is called compliers. Those who take the treatment regardless of the instrument status, $D_{1i} = D_{0i} = 1$, are always-takers, and those who do not take the treatment anyway, $D_{1i} = D_{0i} = 0$, are never-takers. We cannot identify ATE for always-takers and never-takers in general. By the monotonicity assumption of Imbens and Angrist (1994), we exclude defiers who behave in the opposite way with compliers, $D_{1i} = 0$ and $D_{0i} = 1$. Since the compliers are specific to the instrument Z_i , LATE is instrument-specific.

The above setting can be generalized to cases where the number of (excluded) instruments is greater than the number of endogenous variables. The two-stage least squares (2SLS) estimator is widely used to estimate the causal effect in such cases.

Without loss of generality, consider mutually exclusive binary instruments, Z_i^j for $j = 1, \dots, q$. Let D_{zi}^j be i 's potential treatment status when $Z_i^j = z$ where $z = 0, 1$, and $j = 1, \dots, q$. Each instrument identifies a version of LATE because compliers may differ for each Z_i^j . Angrist and Imbens (1995) show that the 2SLS estimator using multiple instruments is consistent for a weighted average of treatment effects for instrument-specific compliers:

$$\rho_a = \sum_{j=1}^q \xi_j \cdot E[Y_{1i} - Y_{0i} | D_{1i}^j > D_{0i}^j], \quad (1.2)$$

where $0 \leq \xi_j \leq 1$ and $\sum_j \xi_j = 1$. Heckman and Vytlacil (2005) extend this result by allowing continuous instruments with covariates. These works provided theoretical foundations to interpret 2SLS point estimates as a weighted average of LATEs, and empirical researchers have done so, either explicitly or implicitly. Examples include Angrist and Chen (2011), Angrist and Evans (1998), Angrist and Krueger (1991), Angrist, Lavy, and Schlosser (2010), Clark and Royer (2013), Dinkelman (2011), Doyle Jr. (2008), Evans and Garthwaite (2012), Evans and Lien (2005), Lochner and Moretti (2004), Thornton (2008), Siminski and Ville (2011), Stephens Jr. and Yang (2014), among many others.

If the 2SLS estimand is a weighted average of more than one LATE, then the commonly conducted over-identifying restrictions test (the J test, hereinafter) would be rejected. A rejection of the J test implies that the postulated moment condition is likely to be misspecified. What is less well known and often overlooked in the literature is that the conventional standard errors are no longer correct under misspecification of the moment condition.¹ This fact has been neglected and the standard errors have been routinely calculated even with small p values of the J test. I derive the asymptotic distribution of 2SLS when the estimand is a weighted average of LATEs, and propose a consistent estimator for the asymptotic variance robust to multiple LATEs. The correct standard error based on the proposed variance estimator can be substantially different from the conventional heteroskedasticity-robust one even for a large sample size, or even for p -values above any usual significance level.

Two recent papers cover similar topics with this one. Kolesár (2013) shows that under treatment effect heterogeneity the 2SLS estimand is a convex combination of

¹The J test can also be rejected due to invalid instruments. Kitagawa (forthcoming) proposed a specification test for instrument validity in this framework.

LATEs while the limited information maximum likelihood (LIML) estimand may not. Angrist and Fernandez-Val (2013) propose an estimand for new subpopulations by reweighting covariate-specific LATEs. However, neither of the two papers considers correct variance estimation of 2SLS.

In the next section, well-known examples of weighted averages of LATEs are replicated, and correct standard errors for 2SLS in those examples are calculated. In Section 3, I show that the postulated moment condition is misspecified when there are more than one LATE. The asymptotic distribution of 2SLS estimators in such a case is derived, and a consistent variance estimator is proposed. Section 4 concludes. The proofs of propositions are collected in the appendix.

2 Weighted Averages of LATEs

How often are researchers interested in a weighted average of LATEs? More common than one might think. In this section, I replicate well-known studies with such interpretations, and show the correct multiple-LATEs-robust standard errors can be substantially different from the reported ones.

First example is Angrist and Krueger (1991), who study the returns to education. Since individuals with higher ability would earn more as well as take more schooling, the relationship between education and earnings cannot be correctly estimated by OLS. The authors avoid endogeneity of education by instrumenting it using quarter of birth (QOB). Individuals who were born in the end of the year enter school at a younger age compared with their classmates. As a result, they are required to take more compulsory schooling before they reach a legal dropout age. Angrist and Krueger estimate the following 2SLS model:

$$\ln W_i = X_i' \beta + \sum_{c=1}^9 Y_{ic} \psi_c + E_i \rho + \epsilon_i, \quad (2.1)$$

$$E_i = X_i' \pi + \sum_{c=1}^9 Y_{ic} \delta_c + \sum_{c=1}^{10} \sum_{j=1}^3 Y_{ic} Q_{ij} \theta_{jc} + u_i, \quad (2.2)$$

where E_i is education, X_i is a vector of covariates including a constant, Y_{ic} is year of birth (YOB), Q_{ij} is QOB, and W_i is weekly wage. If we assume that X_i only contains a constant, then the first stage equation (2.2) is saturated. In this case, the 2SLS

estimand is a weighted average of returns to education where averaging takes place on three different levels. First, for each level of education, it is LATE for those who would have additional schooling because of their QOB. Second, it is averaged over different levels of education because it takes values from 0 to 20. This parameter is also referred to as the average causal response (ACR; Angrist and Imbens, 1995). Lastly, it is averaged over different years of birth because YOB dummies are included in both the first and second stage equations. Since interactions terms between YOB dummies and education are not included in the second stage, it is assumed that the returns to education does not vary over birth years although intercepts may differ. Even if the returns vary, its average is still an interesting and useful estimand because there is no reason that a particular year, e.g. men born in 1930, is more interesting than the cohort of those born in 1930-1939. Therefore it is important to correctly calculate the standard error of the point estimates in this example.

Table 1 shows replication results of Tables IV-VI in Angrist and Krueger along with the multiple-LATEs-robust standard errors (Column MR, in bold). The results for covariates are suppressed. There are a few interesting findings. First, even with large sample sizes, the two standard errors are substantially different. The conventional ones (Column C) are underestimated in all specifications. Second, large p-values do not necessarily mean that the two standard errors are similar. Finally, point estimates averaged over a large set of instruments are more robust. This case is illustrated by Table 3 Column (0), when three QOB dummies are used as only instruments with YOB dummies as covariates. Surprisingly, the returns to education is estimated to be negative. Further inspection reveals that it is a linear combination of three IV estimates, -0.0191 (0.0272) using only the first quarter as an instrument, -1.3167 (5.2517) using the second quarter, 0.2858 (0.1932) using the third quarter, where the numbers in parentheses are conventional IV standard errors. Apparently, imprecisely estimated point estimate with the second QOB is the main reason for the negative point estimate in Column (0). Since the F statistic is much larger than the rule of thumb, 10, it appears that weak instruments are not an issue. The researcher might get around the problem by using a different instrument, but a better alternative is to get a 2SLS estimate based on a larger set of instruments.

Second example is Angrist and Evans (1998) who use the sex of mother's first two children as instruments to estimate the effect of family size on mother's labor supply. The instruments *two-boys* and *two-girls* are based on the fact that American

	Column	ρ	C	MR	p-value of J test	dof
Table IV:	(0)	.0634	.0166	.0167	.3136	2
Men Born	(2)	.0769	.0151	.0170	.1661	29
1920-1929,	(4)	.1310	.0336	.0454	.5359	27
$n = 247, 199$	(6)	.0669	.0152	.0169	.2196	29
	(8)	.1007	.0336	.0474	.3578	27
Table V:	(0)	.1053	.0201	.0204	.1917	2
Men Born	(2)	.0891	.0162	.0176	.6935	29
1930-1939,	(4)	.0760	.0292	.0359	.7110	27
$n = 329, 509$	(6)	.0806	.0165	.0178	.8184	29
	(8)	.0600	.0292	.0349	.8614	27
Table VI	(0)	-.0612	.0259	.0275	.0042	2
Men Born	(2)	.0553	.0138	.0166	.0000	29
1940-1949,	(4)	.0948	.0221	.0277	.0049	27
$n = 486, 926$	(6)	.0393	.0146	.0175	.0000	29
	(8)	.0779	.0238	.0308	.0033	27

Table 1: Comparison of the proposed multiple-LATEs-robust (MR) and the conventional (C) standard errors—Replication of Table IV, V, and VI in Angrist and Krueger (1991). Each column corresponds to different sets of covariates and instruments. Column (0) refers to the case that only the three quarter-of-birth dummies are used as instruments, calculated by the author.

parents tended to go for a third child when their first two children were the same sex. Each of the instruments identifies LATE of those whose fertility was affected by their children’s sex mix, and the two LATEs are not necessarily the same. A specification used by Angrist and Evans is

$$Y_i = X_i' \pi + M_i \rho + \epsilon_i, \quad (2.3)$$

$$M_i = X_i' \pi + TB_i \cdot \theta_1 + TG_i \cdot \theta_2 + u_i, \quad (2.4)$$

where M_i is an indicator for more than 2 children, TB_i and TG_i are indicators for two-boys and two-girls, X_i is a vector of covariates including a constant and an indicator for first boy, and Y_i is an indicator for whether the respondent worked for pay in the Census year. The OLS estimate of ρ is -.167, but it is argued to exaggerate the causal effect of fertility on female labor supply due to selection bias. Using the

instruments one at a time, we get the IV estimates $-.201$ for two-boys and $-.059$ for two-girls instrument. The large (small) reduction in labor supply using the two-boys (two-girls) instrument looks reasonable, but it is difficult to compare it with the OLS estimate because the latter is based on the whole population, while the IV estimates are for some subpopulations. Since the ultimate goal is to estimate the overall effect of having more than two children on mother’s labor supply, one strategy is to calculate an average of the two IV estimates. 2SLS estimand is a particular weighted average where the weights are implicitly calculated based on the relative strength of each instrument.²

Table 2 shows replication results of Table 7 in Angrist and Evans (1998). First of all, individual IV estimates do not give a satisfactory answer whether the OLS estimate is biased or not, while 2SLS estimates do across different specifications. Second, unlike the replication of Angrist and Krueger (1991), the two standard errors are almost the same, even the p values are quite small. Thus, it is not always the case that the correct standard error is much larger than the conventional one. Since they are similar, there is a sizable gain in precision by combining the two instruments, compared with using a single instrument. Lastly, the 2SLS point estimates are weighted averages of the two IV estimates, where the weight for the two-boys instrument is $.38$. Since the weight is completely determined by the first stage, the same weight is used across different dependent variable in the second stage. In this example, the two-boys instrument receives less weight because the first-stage coefficient is smaller, which implies that the absolute size of the compliers is smaller than that of the two-girls instrument. The proposed multiple-LATES-robust standard error can be computed for other weighted averages of LATEs, as long as they can be written as a GMM estimator.

3 Moment condition for 2SLS

In this section, I link the identification of LATEs and estimation of such parameters by moment condition models. I maintain the assumption that the treatment variable

²In this example, 2SLS estimand is not exactly equal to a weighted average of covariate-specific LATEs, because the first stage is not fully saturated. However, Angrist (2001) shows that 2SLS estimates using the twins instruments are almost the same with the one using a fully saturated first stage based on the procedure of Abadie (2003), and Angrist and Pischke (2009) argue that this is likely to hold in practice.

Y_i	Estimator	ρ	C	MR	p-value of J test
Worked for pay	2SLS (both)	-.1128	.0277	.0277	.0129
	IV (two-boys)	-.2011	.0450		–
	IV (two-girls)	-.0591	.0352		–
	OLS	-.1666	.0020		–
Weeks worked	2SLS (both)	-5.164	1.201	1.203	.0711
	IV (two-boys)	-7.944	1.950		–
	IV (two-girls)	-3.473	1.527		–
	OLS	-8.044	.087		–
Hours/week	2SLS (both)	-4.613	1.008	1.010	.0492
	IV (two-boys)	-7.159	1.644		–
	IV (two-girls)	-3.065	1.279		–
	OLS	-6.021	.074		–
Labor income	2SLS (both)	-1321.2	566.4	566.4	.7025
	IV (two-boys)	-1597.8	914.7		–
	IV (two-girls)	-1153.0	721.3		–
	OLS	-3165.4	40.6		–

Table 2: Comparison of the proposed multiple-LATEs-robust (MR) and the conventional (C) standard errors—Replication of Table 7 Columns (4) and (6) in Angrist and Evans (1998). The IV estimators using either the two-boys or two-girls instrument are calculated by the author.

and instruments are binary for simplicity of exposition, but this will be relaxed later in this section. The observed outcome can be written as

$$Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i) = E[Y_{0i}] + D_i\rho_i + \eta_i, \quad (3.1)$$

where $\rho_i = Y_{1i} - Y_{0i}$ and $\eta_i = Y_{0i} - E[Y_{0i}]$. Since the individual treatment effect ρ_i cannot be identified, a version of average treatment effect (ATE) becomes the parameter of interest. Let ρ be the parameter, and α be a nuisance parameter for the intercept. Rewriting (3.1) in a familiar regression notation using α and ρ , we get

$$Y_i = \alpha + D_i\rho + e_i \equiv \mathbf{X}'_i\boldsymbol{\beta} + e_i, \quad (3.2)$$

$$e_i \equiv e_i(\alpha, \rho) = E[Y_{0i}] - \alpha + D_i(\rho_i - \rho) + \eta_i, \quad (3.3)$$

where $\mathbf{X}_i = (1, D_i)'$, $\boldsymbol{\beta} = (\alpha, \rho)'$, and e_i is the population residual. This formulation directly connects identification results under treatment effect heterogeneity and estimation using conventional estimators based on moment conditions. For example, the OLS estimand solves the moment condition

$$0 = E[e_i(\alpha_{ols}, \rho_{ols})] = E[D_i e_i(\alpha_{ols}, \rho_{ols})], \quad (3.4)$$

and this gives $\alpha_{ols} = E[Y_i] - E[D_i]\rho_{ols}$ and $\rho_{ols} = E[\rho_i|D_i = 1] + \textit{selection bias}$, where the selection bias term is defined as $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$. Note that the conventional endogeneity condition of OLS is $E[D_i \eta_i] \neq 0$, because $E[D_i e_i(\alpha_{ols}, \rho_{ols})] = 0$ always holds. Now suppose that there is a binary instrument Z_i^1 which satisfies Conditions 1-3 of Imbens and Angrist (1994). Since it is a valid instrument, it satisfies $E[Z_i^1 \eta_i] = 0$. The moment condition for the IV estimator is

$$0 = E[e_i(\alpha_{IV}^1, \rho_{IV}^1)] = E[Z_i^1 e_i(\alpha_{IV}^1, \rho_{IV}^1)], \quad (3.5)$$

and the solution is given by

$$\alpha_{IV}^1 = E[Y_i] - E[D_i] \cdot \rho_{IV}^1, \quad (3.6)$$

$$\rho_{IV}^1 = \frac{\text{Cov}(Y_i, Z_i^1)}{\text{Cov}(D_i, Z_i^1)} = E[\rho_i | D_{1i}^1 > D_{0i}^1]. \quad (3.7)$$

The last equality holds by the LATE theorem of Imbens and Angrist (1994).

In general, the moment conditions such as (3.4) or (3.5) are called just-identified because the number of parameters is equal to the dimension of the moment condition. Just-identified moment conditions always have a solution under regularity conditions, regardless of whether the structural assumptions hold or not. For example, (3.4) always holds, even if $E[D_i \eta_i] \neq 0$. Also, (3.5) always holds, regardless of whether $E[Z_i^1 \eta_i] = 0$ is true or not. If the structural assumptions are violated, then the estimand may not be the parameter of interest, but it does not affect the asymptotic variance of the estimator because the asymptotic distribution is derived under (3.4) or (3.5).

This conclusion does not hold if there are more instruments than the endogenous parameters. In this case, the moment condition is over-identified and the assumption that there exists a solution to the moment condition may be violated. Suppose that

there are two valid instruments, Z_i^1 and Z_i^2 , such that $E[Z_i^1 \eta_i] = E[Z_i^2 \eta_i] = 0$. If we use each instrument one at a time, we would get ρ_{IV}^1 and ρ_{IV}^2 , where each corresponds to a different LATE. To use both instruments at the same time, the 2SLS procedure involves two stages, where the first stage $D_i = \delta + Z_i^1 \pi_1 + Z_i^2 \pi_2 + u_i$ is estimated by OLS to produce the fitted value \hat{D}_i , and the second stage $Y_i = \alpha + \hat{D}_i \rho + e_i$ is estimated by OLS. Angrist and Imbens (1995) show that the 2SLS estimand is

$$\alpha_0 = E[Y_i] - E[D_i] \cdot \rho_0, \quad (3.8)$$

$$\rho_0 = \xi \cdot \rho_{IV}^1 + (1 - \xi) \cdot \rho_{IV}^2, \quad 0 \leq \xi \leq 1. \quad (3.9)$$

The 2SLS estimator is equivalent to a GMM estimator based on the postulated moment condition

$$0 = E[e_i(\alpha_0, \rho_0)] = E[Z_i^1 e_i(\alpha_0, \rho_0)] = E[Z_i^2 e_i(\alpha_0, \rho_0)], \quad (3.10)$$

for a unique parameter (α_0, ρ_0) . By using the fact that $e_i(\alpha_0, \rho_0) = Y_i - \alpha_0 - D_i \rho_0$, it is straightforward to show that (3.10) does not hold unless $\rho_0 = \rho_{IV}^1 = \rho_{IV}^2$ which implies that the two LATEs are the same. This need not be true under treatment effect heterogeneity. Thus, we conclude that $E[Z_i^1 e_i(\alpha_0, \rho_0)] \neq 0$ and $E[Z_i^2 e_i(\alpha_0, \rho_0)] \neq 0$. This holds if there are more than two instruments with covariates. Thus, the postulated moment condition of 2SLS is misspecified.

Misspecified moment conditions under treatment effect heterogeneity have important implications. First, the J test will reject the null hypothesis of (3.10) asymptotically. It is not surprising that researchers often face a significant J test statistic when multiple instruments are used. If we can rule out the possibility of invalid instruments, $E[Z_i^j \eta_i] \neq 0$ for some j , either by a statistical test such as Kitagawa (forthcoming), or by an economic reasoning, the rejection is due to treatment effect heterogeneity. Thus, conducting the J test has little relevance if heterogeneity is already assumed. Second, the asymptotic variance of 2SLS will be different from the standard one, and the conventional heteroskedasticity-robust variance estimator would be inconsistent. It is surprising that this has been overlooked in the literature. In the following propositions, I derive the asymptotic distribution of 2SLS and propose a consistent variance estimator. The implications suggest that the proposed variance estimator should always be used to calculate the standard error of 2SLS, regardless of the J test

results.

The above result can be generalized to models with covariates, and situations where instruments or a treatment variable can take multiple values. Angrist and Imbens (1995) define 2SLS estimand with covariates, a discrete treatment variable, and multiple instruments when the first stage is fully saturated. Kolesár (2013) extends this by allowing different instruments sets in the first stage. Heckman and Vytlacil (2005) show that IV estimand (including 2SLS) is a function of the marginal treatment effect for continuous instruments. These works characterize the 2SLS estimand in a general setting, but neither the asymptotic distribution nor a consistent variance estimator is provided.

To formally derive the asymptotic distribution, I consider the model (3.2) with covariates. Assume that there are valid instruments $Z_i^1, Z_i^2, \dots, Z_i^q$ such that $E[Z_i^j \eta_i] = 0$ for $j = 1, \dots, q$. Let $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)_{i=1}^n$ be an iid sample, where $\mathbf{X}_i = (\mathbf{W}_i', D_i)'$, $\mathbf{Z}_i = (\mathbf{W}_i', Z_i^1, \dots, Z_i^q)'$, and \mathbf{W}_i be an $l \times 1$ vector of covariates including a constant. The first and second stages are

$$Y_i = \mathbf{W}_i' \boldsymbol{\gamma} + D_i \rho + e_i \equiv \mathbf{X}_i \boldsymbol{\beta} + e_i, \quad (3.11)$$

$$D_i = \mathbf{W}_i' \boldsymbol{\delta} + \mathbf{Z}_i' \boldsymbol{\pi} + u_i. \quad (3.12)$$

I emphasize that e_i should be interpreted as the population residual, similar to (3.3), not as the structural error η_i . In addition, u_i is the projection error. The 2SLS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Y}, \quad (3.13)$$

where $\mathbf{X} \equiv (\mathbf{X}'_1, \dots, \mathbf{X}'_n)'$ is an $n \times (l+1)$ matrix, $\mathbf{Z} \equiv (\mathbf{Z}'_1, \dots, \mathbf{Z}'_n)'$ is an $n \times (l+q)$ matrix, and $\mathbf{Y} \equiv (Y_1, \dots, Y_n)'$ is an $n \times 1$ vector. The following proposition establishes the asymptotic distribution of 2SLS estimators when there are more than one LATE in a general setting.

Proposition 1. *Let $\boldsymbol{\beta}_0 = (\boldsymbol{\gamma}_0', \rho_0)'$ be the 2SLS estimand where $\boldsymbol{\gamma}_0$ satisfies $E[Y_i] = E[\mathbf{W}_i'] \boldsymbol{\gamma}_0 + E[D_i] \rho_0$ and ρ_0 is a linear combination of different LATEs. Let $e_i \equiv Y_i - \mathbf{X}'_i \boldsymbol{\beta}_0$. The asymptotic distribution of 2SLS is*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, H^{-1} \Omega H^{-1}),$$

where $H = E[\mathbf{X}_i \mathbf{Z}'_i] (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} E[\mathbf{Z}_i \mathbf{X}'_i]$, $\Omega = E[\boldsymbol{\psi}_i \boldsymbol{\psi}'_i]$, and

$$\begin{aligned} \boldsymbol{\psi}_i &= E[\mathbf{X}_i \mathbf{Z}'_i] (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} (\mathbf{Z}_i e_i - E[\mathbf{Z}_i e_i]) + (\mathbf{X}_i \mathbf{Z}'_i - E[\mathbf{X}_i \mathbf{Z}'_i]) (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} E[\mathbf{Z}_i e_i] \\ &\quad + E[\mathbf{X}_i \mathbf{Z}'_i] (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} (E[\mathbf{Z}_i \mathbf{Z}'_i] - \mathbf{Z}_i \mathbf{Z}'_i) (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} E[\mathbf{Z}_i e_i]. \end{aligned}$$

The next proposition proposes a consistent estimator for the asymptotic variance matrix of 2SLS robust to multiple-LATEs.

Proposition 2. *A multiple-LATEs-robust asymptotic variance estimator for 2SLS is given by*

$$\hat{\boldsymbol{\Sigma}}_{MR} = n \cdot \left(\mathbf{X}'\mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \left(\sum_i \hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}'_i \right) \left(\mathbf{X}'\mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \quad (3.14)$$

where

$$\begin{aligned} \hat{\boldsymbol{\psi}}_i &= \frac{1}{n} \mathbf{X}'\mathbf{Z} \left(\frac{1}{n} \mathbf{Z}'\mathbf{Z} \right)^{-1} \left(\mathbf{Z}_i \hat{e}_i - \frac{1}{n} \mathbf{Z}'\hat{\mathbf{e}} \right) \\ &\quad + \left(\mathbf{X}_i \mathbf{Z}'_i - \frac{1}{n} \mathbf{X}'\mathbf{Z} \right) \left(\frac{1}{n} \mathbf{Z}'\mathbf{Z} \right)^{-1} \frac{1}{n} \mathbf{Z}'\hat{\mathbf{e}} \\ &\quad + \frac{1}{n} \mathbf{X}'\mathbf{Z} \left(\frac{1}{n} \mathbf{Z}'\mathbf{Z} \right)^{-1} \left(\frac{1}{n} \mathbf{Z}'\mathbf{Z} - \mathbf{Z}_i \mathbf{Z}'_i \right) \left(\frac{1}{n} \mathbf{Z}'\mathbf{Z} \right)^{-1} \frac{1}{n} \mathbf{Z}'\hat{\mathbf{e}}, \end{aligned} \quad (3.15)$$

$\hat{e}_i = Y_i - \mathbf{X}'_i \hat{\boldsymbol{\beta}}$, and $\hat{\mathbf{e}} = (\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n)'$.

The formula of $\hat{\boldsymbol{\Sigma}}_{MR}$ is different from that of the conventional heteroskedasticity-robust variance estimator:

$$\hat{\boldsymbol{\Sigma}}_C = n \cdot \left(\mathbf{X}'\mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \left(\sum_i \mathbf{Z}_i \mathbf{Z}'_i \hat{e}_i^2 \right) \left(\mathbf{X}'\mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1}. \quad (3.16)$$

Under homogeneous treatment effect, both $\hat{\boldsymbol{\Sigma}}_{MR}$ and $\hat{\boldsymbol{\Sigma}}_C$ have the same probability limit, but they are generally different in finite sample. $\hat{\boldsymbol{\Sigma}}_{MR}$ is consistent for the true asymptotic variance matrix even when the postulated moment condition is misspecified, and thus can be used regardless of whether there is one or more than one LATE. In contrast, $\hat{\boldsymbol{\Sigma}}_C$ is consistent only if the underlying LATEs are the same. This is also true for the standard errors based on $\hat{\boldsymbol{\Sigma}}_{MR}$ and $\hat{\boldsymbol{\Sigma}}_C$.

When there is a single endogenous variable without covariates, Proposition 1 coincides with the result in the proof of Theorem 3 of Imbens and Angrist (1994) when the first stage is known but needs to be estimated.³ Their derivation is based on the stacked moment condition that consists of first-order conditions of the first and second stages. This is a special case of general two-step estimators of Newey and McFadden (1994). The derivation of Imbens and Angrist uses the condition that the population fitted value of the endogenous variable is uncorrelated with e_i , where e_i is defined in Proposition 1. For example, the condition is $E[(\delta + \pi_1 Z_i^1 + \pi_2 Z_i^2)e_i] = 0$ for a two instruments case, which does not necessarily imply $E[Z_i^1 e_i] = E[Z_i^2 e_i] = 0$. This makes their asymptotic variance and its estimator robust to violations of the underlying moment condition, $E[Z_i^1 e_i] = E[Z_i^2 e_i] = 0$. Thus, they coincide with Σ_{MR} and $\hat{\Sigma}_{MR}$. Even in such cases, however, their formula has not been used in practice. Statistics softwares such as Stata do not estimate their asymptotic variance, but the standard GMM one assuming correct specification. This results in wrong standard errors. The main contribution of this paper is to observe that 2SLS using multiple instruments under treatment effect heterogeneity is a special case of misspecified GMM of Hall and Inoue (2003). Specifically, Proposition 1 is a special case of their Theorem 2 in the context of treatment effect heterogeneity.

Since the 2SLS estimator is a special case of a GMM estimator using $(\mathbf{Z}'\mathbf{Z})^{-1}$ as a weight matrix, we may consider an alternative GMM estimator based on another weight matrix. This will lead to a different weighted average of LATEs, which may be more appealing than the conventional 2SLS estimand. Let $E[\mathbf{L}_i \mathbf{L}_i']$ be an alternative symmetric positive definite matrix where \mathbf{L}_i is an $(l+q) \times 1$ vector, and let $(\mathbf{L}'\mathbf{L})^{-1}$ be the sample weight matrix, where \mathbf{L} is an $n \times (l+q)$ matrix. The alternative estimator based on the same moment condition but a different weight matrix is given by

$$\tilde{\beta} = (\mathbf{X}'\mathbf{Z}(\mathbf{L}'\mathbf{L})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{L}'\mathbf{L})^{-1}\mathbf{Z}'\mathbf{Y}. \quad (3.17)$$

Let β_a be the probability limit of the estimator. The asymptotic distribution of $\sqrt{n}(\tilde{\beta} - \beta_a)$ and the variance estimator can be obtained by slight modifications of Propositions 1 and 2. In particular, replace $\mathbf{Z}_i \mathbf{Z}_i'$ with $\mathbf{L}_i \mathbf{L}_i'$, $\mathbf{Z}'\mathbf{Z}$ with $\mathbf{L}'\mathbf{L}$, and \hat{e}_i

³There are typos in the proof of Theorem 3 of Imbens and Angrist (1994). Their matrix Δ should read $\Delta = \begin{pmatrix} E[\psi(Z, D, \theta) \cdot \psi(Z, D, \theta)'] & E[\varepsilon \cdot \psi(Z, D, \theta)] & E[g(Z) \cdot \varepsilon \cdot \psi(Z, D, \theta)] \\ E[\varepsilon \cdot \psi(Z, D, \theta)]' & E[\varepsilon^2] & E[g(Z) \cdot \varepsilon^2] \\ E[g(Z) \cdot \varepsilon \cdot \psi(Z, D, \theta)]' & E[g(Z) \cdot \varepsilon^2] & E[g^2(Z) \cdot \varepsilon^2] \end{pmatrix}$.

with $\tilde{e}_i = Y_i - \mathbf{X}_i' \tilde{\boldsymbol{\beta}}$, whenever they appear.

Remark 1 (Using the propensity score as an instrument). When D_i is binary, Heckman and Vytlacil (2005) show that the propensity score, $P(D_i = 1 | \mathbf{W}_i, \mathbf{Z}_i)$, has a few desirable properties when used as an instrument. With the same set of instruments and covariates in the 2SLS first stage, one can estimate the (nonlinear) propensity score. The resulting IV estimate⁴ would differ from 2SLS point estimate, but they are both valid based on different weighting. One may wonder if the proposed variance estimator can be used for the IV estimator in this case. Since it is not a linear GMM estimator, the proposed formula cannot be used to calculate the standard error. Instead, the formula for two-step estimators of Newey and McFadden (1994) can be used. The formula for a logit first stage with covariates is given in the appendix. On the other hand, the standard 2SLS would yield the same point estimate with the IV estimator using the propensity score as an instrument when the first stage is fully saturated, because the first stage consistently estimates the propensity score. In this case, the standard error can be calculated using Proposition 2.

Remark 2 (Invalid Instruments). The proposed multiple-LATEs-robust variance estimator $\hat{\boldsymbol{\Sigma}}_{MR}$ is also robust to invalid instruments, i.e., instruments correlated with the error term. Consider a linear model $Y_i = \mathbf{X}_i' \boldsymbol{\beta}_0 + e_i$ where \mathbf{X}_i is a $(k + p) \times 1$ vector of regressors. Among $k + p$ regressors, p are endogenous, i.e. $E[\mathbf{X}_i e_i] \neq 0$. If a $k + q$ vector of instruments \mathbf{Z}_i is available such that $E[\mathbf{Z}_i e_i] = 0$ and $q \geq p$, then $\boldsymbol{\beta}_0$ can be consistently estimated by 2SLS or GMM. If any of the instruments is invalid, then $E[\mathbf{Z}_i e_i] \neq 0$ and $\boldsymbol{\beta}_0$ may not be consistently estimated. Instead, a pseudo-true value that minimizes the corresponding GMM criterion is estimated.⁵ Since the moment condition does not hold, the model is misspecified. There are two types of misspecification: (i) fixed or global misspecification such that $E[\mathbf{Z}_i e_i] = \boldsymbol{\delta}$ where $\boldsymbol{\delta}$ is a constant vector containing at least one non-zero component, and (ii) local misspecification such that $E[\mathbf{Z}_i e_i] = n^{-r} \boldsymbol{\delta}$ for some $r > 0$. A particular choice of $r = 1/2$ has been used to analyse the asymptotic behavior of 2SLS estimators with invalid instruments by Hahn and Hausman (2005), Bravo (2010), Berkowitz, Caner,

⁴If a nonlinear model such as probit or logit is used to estimate the propensity score in the first stage, then the usual 2SLS procedure cannot be directly applied because the first stage residuals are not uncorrelated with fitted values and covariates. This is often called a *forbidden regression*. Instead, one can use the estimated propensity score as the instrument and calculate the IV estimator.

⁵The 2SLS estimand $\boldsymbol{\beta}_0$ in Proposition 1 is an example of such pseudo-true values.

and Fang (2008, 2012), Otsu (2011), Guggenberger (2012), and DiTraglia (2015). Under either fixed or local misspecification, $\hat{\Sigma}_{MR}$ in Proposition 2 is consistent for the true asymptotic variance. However, the conventional variance estimator $\hat{\Sigma}_C$ is inconsistent under fixed misspecification. Under local misspecification, $\hat{\Sigma}_C$ is consistent but the rate of convergence is negatively affected.

Remark 3 (Bootstrap). Bootstrapping can be used to get more accurate t tests and confidence intervals (CI's) based on $\hat{\beta}$, in terms of having smaller errors in the rejection probabilities or coverage probabilities. This is called asymptotic refinements of the bootstrap. Since the model is over-identified and possibly misspecified, and 2SLS is a special case of GMM, the misspecification-robust bootstrap for GMM of Lee (2014) can be used. In contrast, the conventional bootstrap methods for over-identified GMM of Hall and Horowitz (1996), Brown and Newey (2002), and Andrews (2002) assume correctly specified moment conditions. Since this implies the constant treatment effect, they achieve neither asymptotic refinements nor consistency in this context. Suppose one wants to test $H_0 : \beta_m = \beta_{0,m}$ or to construct a CI for $\beta_{0,m}$ where $\beta_{0,m}$ is the m th element of β_0 . The misspecification-robust bootstrap critical values for t tests and CI's are calculated from the simulated distribution of the bootstrap t statistic

$$T_n^* = \frac{\hat{\beta}_m^* - \hat{\beta}_m}{\sqrt{\hat{\Sigma}_{MR,m}^*/n}}$$

where $\hat{\beta}_m^*$ and $\hat{\beta}_m$ are the m th elements of $\hat{\beta}^*$ and $\hat{\beta}$, respectively, $\hat{\Sigma}_{MR,m}^*$ is the m th diagonal element of $\hat{\Sigma}_{MR}^*$, and $\hat{\beta}^*$ and $\hat{\Sigma}_{MR}^*$ are the bootstrap versions of $\hat{\beta}$ and $\hat{\Sigma}_{MR}$ based on the same formula using the bootstrap sample rather than the original sample.

4 Conclusion

Two-stage least squares (2SLS) estimators are widely used in practice. When heterogeneity is present in treatment effects, 2SLS point estimates can be interpreted as a weighted average of the local average treatment effects (LATE). I show that the conventional standard errors, typically generated by econometric softwares such as Stata, are incorrect in this case. The over-identifying restrictions test is often used to test the presence of heterogeneity, but it is not useful in this context because it can also

reject due to invalid instruments. I provide a simple standard error formula for 2SLS which is correct regardless of whether there are multiple LATEs or not. In addition, this standard error is robust to invalid instruments, and can be used for bootstrapping to achieve asymptotic refinements under treatment effect heterogeneity.

Appendix

A Proofs of Propositions

Proposition 1

Proof. Let $\mathbf{e} \equiv (e_1, \dots, e_n)'$ be an $n \times 1$ vector where $e_i \equiv Y_i - \mathbf{X}'_i \boldsymbol{\beta}_a$. Evaluated at $\boldsymbol{\beta}_a$, the moment condition does not hold:

$$E[\mathbf{Z}_i(Y_i - \mathbf{X}'_i \boldsymbol{\beta}_a)] \equiv E[\mathbf{Z}_i e_i] \neq 0, \quad (\text{A.1})$$

if there are more than one LATE. This can be shown by the following argument. For simplicity, assume that we have two instruments, Z_i^1 and Z_i^2 , such that each instrument satisfies regularity conditions for identifying the instrument-specific LATE. Let ρ^j be the LATE with respect to Z_i^j and $\boldsymbol{\beta}^j \equiv (\boldsymbol{\gamma}^{j'}, \rho^j)'$ be the parameter vector for $j = 1, 2$. By assumption, $\boldsymbol{\beta}^1 \neq \boldsymbol{\beta}^2$. If we use each instrument at a time, $E[Z_i^1(Y_i - \mathbf{X}'_i \boldsymbol{\beta}^1)] = E[Z_i^2(Y_i - \mathbf{X}'_i \boldsymbol{\beta}^2)] = 0$. Now assume $E[\mathbf{Z}_i(Y_i - \mathbf{X}'_i \boldsymbol{\beta}_a)] = 0$ holds. Then $E[Z_i^1(Y_i - \mathbf{X}'_i \boldsymbol{\beta}_a)] = E[Z_i^2(Y_i - \mathbf{X}'_i \boldsymbol{\beta}_a)] = 0$, but this implies $\boldsymbol{\beta}_a = \boldsymbol{\beta}^1 = \boldsymbol{\beta}^2$. This contradicts to the assumption. Thus, (A.1) holds.

From the first-order condition of GMM, we substitute $\mathbf{X}\boldsymbol{\beta}_a + \mathbf{e}$ for \mathbf{Y} , rearrange

terms, and multiply \sqrt{n} to have

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_a) &= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\sqrt{n}\mathbf{Z}'\mathbf{e}, & (\text{A.2}) \\
&= \left(\frac{1}{n}\mathbf{X}'\mathbf{Z} \left(\frac{1}{n}\mathbf{Z}'\mathbf{Z} \right)^{-1} \frac{1}{n}\mathbf{Z}'\mathbf{X} \right)^{-1} \times \\
&\quad \left\{ \frac{1}{n}\mathbf{X}'\mathbf{Z} \left(\frac{1}{n}\mathbf{Z}'\mathbf{Z} \right)^{-1} \sqrt{n} \left(\frac{1}{n}\mathbf{Z}'\mathbf{e} - E[\mathbf{Z}_i e_i] \right) \right. \\
&\quad + \sqrt{n} \left(\frac{1}{n}\mathbf{X}'\mathbf{Z} - E[\mathbf{X}_i \mathbf{Z}'_i] \right) \left(\frac{1}{n}\mathbf{Z}'\mathbf{Z} \right)^{-1} E[\mathbf{Z}_i e_i] \\
&\quad \left. + E[\mathbf{X}_i \mathbf{Z}'_i] \sqrt{n} \left(\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z} \right)^{-1} - (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} \right) E[\mathbf{Z}_i e_i] \right\}.
\end{aligned}$$

The second equality holds because the population first-order condition of GMM holds regardless of misspecification, i.e., $0 = E[\mathbf{X}_i \mathbf{Z}'_i] E[\mathbf{Z}_i \mathbf{Z}_i']^{-1} E[\mathbf{Z}_i e_i]$. The expression (A.2) is different from the standard one because $E[\mathbf{Z}_i e_i] \neq 0$. As a result, the asymptotic variance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_a)$ includes additional terms, which are assumed to be zero in the standard asymptotic variance matrix of 2SLS. We use the fact that

$$\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z} \right)^{-1} - E[\mathbf{Z}_i \mathbf{Z}'_i]^{-1} = (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} \left(E[\mathbf{Z}_i \mathbf{Z}'_i] - \frac{1}{n}\mathbf{Z}'\mathbf{Z} \right) \left(\frac{1}{n}\mathbf{Z}'\mathbf{Z} \right)^{-1}, \quad (\text{A.3})$$

and take the limit of the right-hand-side of (A.2). By the weak law of large numbers (WLLN), the continuous mapping theorem (CMT), and the central limit theorem (CLT),

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_a) \xrightarrow{d} H^{-1} \cdot N(\mathbf{0}, \Omega). \quad (\text{A.4})$$

Q.E.D.

Proposition 2

Proof. Since $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_a$, by WLLN and CMT, $n^{-1} \sum_i \hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}_i'$ is consistent for Ω . By using WLLN and CMT again, $\hat{\boldsymbol{\Sigma}}_{MR}$ is consistent for $H^{-1}\Omega H^{-1}$. *Q.E.D.*

B Asymptotic variance when the logit model is used for the first stage

Let W_i be a covariate vector including a constant. Assume the logit model for the propensity score:

$$P(D_i = 1|W_i, Z_i) = \frac{1}{1 + \exp(-W_i'\delta_0 - Z_i'\pi_0)}. \quad (\text{B.1})$$

Then the log-likelihood function is

$$\mathcal{L}(\delta, \pi) = -\sum_{i=1}^n (1 - D_i)(W_i'\delta + Z_i'\pi) - \sum_{i=1}^n \ln(1 + \exp(-W_i'\delta - Z_i'\pi)). \quad (\text{B.2})$$

The first-order condition of the first stage is

$$0 = n^{-1} \sum_{i=1}^n \begin{pmatrix} W_i \\ Z_i \end{pmatrix} \hat{u}_i \quad (\text{B.3})$$

where

$$u_i(\delta, \pi) = -(1 - D_i) + \frac{\exp(-W_i'\delta - Z_i'\pi)}{1 + \exp(-W_i'\delta - Z_i'\pi)} \quad (\text{B.4})$$

and $\hat{u}_i = u_i(\hat{\delta}, \hat{\pi})$. For the second stage, the first-order condition is

$$0 = n^{-1} \sum_{i=1}^n \begin{pmatrix} W_i \\ \frac{1}{1 + \exp(-W_i'\hat{\delta} - Z_i'\hat{\pi})} \end{pmatrix} \hat{e}_i \quad (\text{B.5})$$

where

$$e_i(\gamma, \rho) = Y_i - W_i'\gamma - D_i'\rho \quad (\text{B.6})$$

and $\hat{e}_i = e_i(\hat{\gamma}, \hat{\rho})$. Now consider a stacked moment function

$$h_i(\beta) = \begin{pmatrix} W_i u_i(\delta, \pi) \\ Z_i u_i(\delta, \pi) \\ W_i e_i(\gamma, \rho) \\ \frac{1}{1 + \exp(-W_i'\delta - Z_i'\pi)} e_i(\gamma, \rho) \end{pmatrix}, \quad (\text{B.7})$$

where $\beta = (\delta, \pi, \gamma, \rho)'$. This forms a just-identified moment condition model. Let $\hat{\beta} = (\hat{\delta}, \hat{\pi}, \hat{\gamma}, \hat{\rho})'$ and $\beta_0 = (\delta_0, \pi_0, \gamma_0, \rho_0)'$ be the probability limit. Using standard asymptotic theory for just-identified GMM, the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$ is $N(0, V)$ where $V = \Gamma^{-1}\Delta(\Gamma')^{-1}$, $\Gamma = E(\partial/\partial\beta')h_i(\beta_0)$, and $\Delta = E[h_i(\beta_0)h_i(\beta_0)']$. A consistent estimator of V can be obtained by replacing the population moments with the sample moments: $\hat{V} = \hat{\Gamma}^{-1}\hat{\Delta}(\hat{\Gamma}')^{-1}$ where $\hat{\Gamma} = n^{-1}\sum_{i=1}^n(\partial/\partial\beta')h_i(\hat{\beta})$ and $\hat{\Delta} = n^{-1}\sum_{i=1}^n h_i(\hat{\beta})h_i(\hat{\beta})'$. The correct standard errors for $\hat{\beta}$ can be obtained by taking the square roots of the diagonal elements of \hat{V} divided by n .

References

- Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2), 231-263.
- Andrews, D. W. (2002). Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators. *Econometrica*, 70(1), 119-162.
- Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors. *Journal of Business and Economic Statistics*, 19(1).
- Angrist, J. D. and Chen, S. H. (2011). Schooling and the Vietnam-Era GI Bill: Evidence from the Draft Lottery. *American Economic Journal: Applied Economics*, 96-118.
- Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review* 88(3), 450-477.
- Angrist, J. D. and Fernandez-Val, I. (2013, May). Extrapolating: External Validity and. In *Advances in Economics and Econometrics: Tenth World Congress (Vol. 3, p. 401)*. Cambridge University Press.
- Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430), 431-442.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444-455.
- Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *Quarterly Journal of Economics*, 106(4), 979-1014.

- Angrist, J., Lavy, V., and Schlosser, A. (2010). Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics*, 28(4), 773-824.
- Angrist, J. and Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton University Press.
- Berkowitz, D., Caner, M., and Fang, Y. (2008). Are nearly exogenous instruments reliable? *Economics Letters* 101 (1), 20-23.
- Berkowitz, D., Caner, M., and Fang, Y. (2012). The validity of instruments revisited. *Journal of Econometrics* 166 (2), 255-266.
- Bravo, F. (2010). Efficient M-estimators with auxiliary information. *Journal of Statistical Planning and Inference* 140 (11), 3326-3342.
- Brown, B. W. and Newey, W. K. (2002). Generalized method of moments, efficient bootstrapping, and improved inference. *Journal of Business and Economic Statistics*, 20(4), 507-517.
- Clark, D. and Royer, H. (2013). The effect of education on adult mortality and health: Evidence from Britain. *American Economic Review*, 103(6), 2087-2120.
- Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from South Africa. *American Economic Review*, 3078-3108.
- DiTraglia, F. J. (2015). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. Working Paper. University of Pennsylvania.
- Doyle Jr, J. J. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of Political Economy*, 116(4), 746-770.
- Evans, W. N., and Garthwaite, C. (2012). Estimating heterogeneity in the benefits of medical treatment intensity. *Review of Economics and Statistics*, 94(3), 635-649.
- Evans, W. N. and Lien, D. S. (2005). The benefits of prenatal care: evidence from the PAT bus strike. *Journal of Econometrics* 125(1), 207-239.
- Guggenberger, P. (2012). On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econometric Theory* 28 (2), 387-421.
- Hahn, J. and Hausman, J. (2005). Estimation with valid and invalid instruments. *Annales d'Économie et de Statistique*, 25-57.

- Hall, A. R. and Inoue, A. (2003). The large sample behavior of the generalized method of moments estimator in misspecified models. *Journal of Econometrics* 114 (2), 361-394.
- Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica*, 64(4), 891-916.
- Heckman, J. J. and Vytlacil, E. (2005). Structural Equations, treatment Effects, and econometric policy evaluation. *Econometrica*, 73(3), 669-738.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* 62(2), 467-475.
- Kitagawa, T. (forthcoming). A Test for Instrument Validity. *Econometrica*.
- Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. unpublished manuscript. Cowles Foundation.
- Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4, 2111-2245.
- Lee, S. (2014). Asymptotic refinements of a misspecification-robust bootstrap for generalized method of moments estimators. *Journal of Econometrics*, 178(3), 398-413.
- Lochner, L. and Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, 94(1), 155-189.
- Otsu, T. (2011). Moderate deviations of generalized method of moments and empirical likelihood estimators. *Journal of Multivariate Analysis* 102 (8), 1203-1216.
- Siminski, P. and Ville, S. (2011). Long-run mortality effects of Vietnam-era army service: Evidence from Australia's conscription lotteries. *American Economic Review*, 345-349.
- Stephens, M. Jr. and Yang, D-Y. (2014). Compulsory education and the benefits of schooling. *American Economic Review*, 104(6), 1777-92.
- Thornton, R. L. (2008). The demand for, and impact of, learning HIV Status. *American Economic Review*, 1829-1863.