Old School IV

Master Joshway

ASSA Continuing Education: January 2020

Instruments Ready?

Organizing IV

I tell the IV story in two iterations, first with constant effects, then with heterogeneous potential outcomes.

- The constant effects framework focuses on selection bias and essential IV mechanics
- Many reasons to instrument ... here's one:
  - The regression we want is long, say:

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + A_i'\gamma + v_i = \alpha + \rho \mathrm{S}_i + \eta_i \tag{1}$$

    Since S_i and $\eta_i = A_i'\gamma + v_i$ are correlated,

$$\frac{Cov(\mathrm{Y}_i, \mathrm{S}_i)}{V(\mathrm{S}_i)} \neq \rho$$

  - The short regression suffers from "ability bias"
- *IV recovers long-regression ρ without observing A_i*
- Why go long? Must be a causal story!

IV Goes Long (in pursuit of causal effects)

- Potential earnings modeled as a function of ability:

$$\mathrm{Y}_{0i} = \alpha + \eta_i = \alpha + A_i'\gamma + v_i,$$

  where we're happy to assume $E[v_i \mathrm{S}_i] = 0$
  - Moving from $s-1$ to s years of schooling yields constant returns:

$$\mathrm{Y}_{s,i} - \mathrm{Y}_{s-1,i} = \rho,$$

    making eq. (1) into a causal model
- A valid instrument, Z_i, is:
  1. correlated with S_i
  2. uncorrelated with $\eta_i = A_i'\gamma + v_i$ and hence with Y_{0i}
- Z_i is *excluded* from the causal model of interest
- Given these assumptions, we have:

$$\rho = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)} = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)} = \frac{\text{"}RF\text{"}}{\text{"}1st\text{"}} \tag{2}$$

Abraham (Wald) Meets Jacob (Bernoulli)

- Repeat our long regression, equation (1):

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + A_i'\gamma + v_i = \alpha + \rho \mathrm{S}_i + \eta_i$$

- By linear CEF, the RF for a Bernoulli (dummy) instrument, Z_i, is

$$\frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{V(\mathrm{Z}_i)} = E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0],$$

  with an analogous formula for $\frac{Cov(\mathrm{S}_i,\mathrm{Z}_i)}{V(\mathrm{Z}_i)}$. This shows:

$$\rho = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)} = \frac{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]}{E[\mathrm{S}_i|\mathrm{Z}_i = 1] - E[\mathrm{S}_i|\mathrm{Z}_i = 0]} \tag{3}$$

- A direct route uses $E[\eta_i|\mathrm{Z}_i] = 0$:

$$E[\mathrm{Y}_i|\mathrm{Z}_i] = \alpha + \rho E[\mathrm{S}_i|\mathrm{Z}_i] \tag{4}$$

  Solving (4) for ρ yields (3)

Angrist and Krueger (1991): Compulsory IV

- Children born in late-quarters start school younger, so are kept in school longer by birthday-based compulsory schooling laws
- There's a powerful first stage supporting this
  - Late-quarter births have more years of schooling
  - This is driven by high school and not college, consistent with the AK-91 CSL story
- Mean schooling and wages by YOB/QOB appear in **Figure 4.1.1**
- The IV *estimator* is the sample analog of (2); with a dummy instrument this becomes the sample analog of (3)
- AK-91 Wald IV for the economic returns to schooling compares average schooling and earnings for early- and later-quarter births
  - The instrument here is $Z_i = 1[QOB_i = 1]$

TABLE III

PANEL A: WALD ESTIMATES FOR 1970 CENSUS—MEN BORN 1920–1929[a]

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) − (2)
ln (wkly. wage)	5.1484	5.1574	−0.00898
			(0.00301)
Education	11.3996	11.5252	−0.1256
			(0.0155)
Wald est. of return to education			0.0715
			(0.0219)
OLS return to education[b]			0.0801
			(0.0004)

Panel B: Wald Estimates for 1980 Census—Men Born 1930–1939

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) − (2)
ln (wkly. wage)	5.8916	5.9027	−0.01110
			(0.00274)
Education	12.6881	12.7969	−0.1088
			(0.0132)
Wald est. of return to education			0.1020
			(0.0239)
OLS return to education			0.0709
			(0.0003)

a. The sample size is 247,199 in Panel A, and 327,509 in Panel B. Each sample consists of males born in the United States who had positive earnings in the year preceding the survey. The 1980 Census sample is drawn from the 5 percent sample, and the 1970 Census sample is from the State, County, and Neighborhoods 1 percent samples.

b. The OLS return to education was estimated from a bivariate regression of log weekly earnings on years of education.

Two-Stage Least Squares (2SLS)

- We *do* IV by doing 2SLS
- This accommodates covariates (controls, X_i) and multiple instruments (dummy or otherwise):

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho \mathrm{S}_i + \eta_i, \tag{5}$$

The first stage and reduced form are

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_i' \pi_{10} + \pi_{11}' \mathrm{Z}_i + \xi_{1i} = \hat{s}_i + \xi_{1i} && (6) \\ \mathrm{Y}_i &= \mathrm{X}_i' \pi_{20} + \pi_{21}' \mathrm{Z}_i + \xi_{2i} && (7) \end{aligned}$$

- The 2SLS "second stage" is obtained by substituting (6) into (5):

$$\begin{aligned} \mathrm{Y}_i &= \alpha' \mathrm{X}_i + \rho[\mathrm{X}_i' \pi_{10} + \pi_{11}' \mathrm{Z}_i] + \rho \xi_{1i} + \eta_i && (8) \\ &= \alpha' \mathrm{X}_i + \rho \hat{s}_i + \rho \xi_{1i} + \eta_i \\ &= \alpha' \mathrm{X}_i + \rho \hat{s}_i + \xi_{2i} \end{aligned}$$

2SLS Notes

- 2SLS subs $\hat{s}_i$ for S_i in (5):

$$Y_i = \alpha' X_i + \rho \hat{s}_i + [\eta_i + \rho \xi_{1i}], \tag{9}$$

  - Because $\hat{s}_i$ and ξ_{2i} are uncorrelated, OLS estimation of (9) identifies ρ
  - In practice, let Stata *ivregress* do it

- Likewise, we get the RF this way

$$\begin{aligned} Y_i &= \alpha' X_i + \rho[X_i' \pi_{10} + \pi_{11}' Z_i] + \rho \xi_{1i} + \eta_i \\ &= X_i'[\alpha + \rho \pi_{10}] + \rho \pi_{11}' Z_i + [\rho \xi_{1i} + \eta_i] \\ &= X_i' \pi_{20} + \pi_{21}' Z_i + \xi_{2i} \end{aligned} \tag{10}$$

- 2SLS implicitly computes the ratio of RF to 1st for each IV:

$$\frac{\pi_{21}}{\pi_{11}} = \rho$$

  In old-school SEMs, the sample analog of this ratio is an *Indirect Least Squares* estimator of ρ

2SLS in AK-91

- We now see that it's the *QOB* $\times$ *YOB first stage* and *reduced form* that are plotted in **Figure 4.1.1**
- The corresponding 2SLS estimates appear in **Table 4.1.1**
- 2SLS matches the QOB earnings pattern (RF) to the QOB pattern in schooling (first stage):

$$\pi_{21} = \rho \pi_{11}$$

  The key p-p-p-pattern here is ...

- Covariates include year-of-birth and state-of-birth dummies, as well as linear and quadratic functions of age in quarters

2SLS is a many-splendored thing

- 2SLS is IV where the instrument is $\hat{s}_i^*$, the residual from a regression of $\hat{s}_i$ on X_i:

$$\frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{V(\hat{s}_i^*)} = \frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{Cov(\mathrm{S}_i, \hat{s}_i^*)}$$

- One-instrument 2SLS is IV where the instrument is $\tilde{z}_i$, the residual from a regression of Z_i on the covs, X_i:

$$\frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{V(\hat{s}_i^*)} = \frac{Cov(\mathrm{Y}_i, \tilde{z}_i)}{Cov(\mathrm{S}_i, \tilde{z}_i)}$$

- One-instrument 2SLS is ILS:

$$\begin{aligned} \frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{V(\hat{s}_i^*)} &= \frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{Cov(\mathrm{S}_i, \hat{s}_i^*)} \\ &= \frac{Cov(\mathrm{Y}_i, \tilde{z}_i)}{Cov(\mathrm{S}_i, \tilde{z}_i)} = \frac{\pi_{21}}{\pi_{11}} \end{aligned}$$

- Over-identified 2SLS is a weighted average of these just-identified IV=ILS estimates (MHE 4.5.1)

2SLS Mistakes

- 2SLS . . . so simple a fool can do it . . .
  - and many do!
- What can go wrong?

- As explained in MHE 4.6.1, three mistakes stubbornly persist:
  - Manual 2SLS
  - Covariate ambivalence
  - Forbidden regressions (from the left and the right)

- These are the bitter fruit of attempts to "improve" upon orthodox 2SLS protocols

- 2SLS is already awesome: let Stata do it!

Group Work

Wald Serves in **Vietnam**

- Key variables

$$\begin{aligned} \mathrm{Z}_i &= \text{randomly assigned draft-eligibility in 1970-72 draft lotteries} \\ \mathrm{D}_i &= \text{a dummy indicating Vietnam-era veterans} \\ \mathrm{Y}_i &= \text{earnings after service} \end{aligned}$$

- The causal effect of Vietnam-era military service is the draft-eligibility RF divided by the draft-eligibility first stage
  - D_i is also a dummy, so the first stage is a diff in probs:

$$\begin{aligned} \frac{Cov(\mathrm{D}_i, \mathrm{Z}_i)}{V(\mathrm{Z}_i)} &= E[\mathrm{D}_i|\mathrm{Z}_i = 1] - E[\mathrm{D}_i|\mathrm{Z}_i = 0] \\ &= P[\mathrm{D}_i = 1|\mathrm{Z}_i = 1] - P[\mathrm{D}_i = 1|\mathrm{Z}_i = 0] \end{aligned}$$

- **Angrist (1990), Figures 1-2 and MHE Table 4.1.3**
- Updated: **Angrist, Chen, and Song (2011)**

Multiple groups and 2SLS

- More to the draft lottery than draft-eligibility: **Angrist and Chen (2011), Figure 1**
- Let $\mathrm{R}_i = j \in \{1, ..., J\}$ denote lottery numbers. Draft-eligibility Wald uses $1[\mathrm{R}_i < 195]$ as an instrument in a just-identified setup
- Using fine-grained info on R_i, we have

$$E[\mathrm{Y}_i|\mathrm{R}_i] = \alpha + \rho P[\mathrm{D}_i = 1|\mathrm{R}_i], \tag{11}$$

since $P[\mathrm{D}_i = 1|\mathrm{R}_i] = E[\mathrm{D}_i|\mathrm{R}_i]$. So we can estimate ρ by fitting:

$$\bar{y}_j = \alpha + \rho \hat{p}_j + \bar{\eta}_j; \; j = 1, ..., J \tag{12}$$

- Efficient GLS for this grouped constant-effects linear model is weighted least squares, weighted by $V(\bar{\eta}_j)$
  - $V(\bar{\eta}_j) = \frac{\sigma_\eta^2}{n_j}$ under homoskedasticity

Visual IV, Grouping, and GLS

- Equation (12) in action: **Angrist (1990), Figure 3**.
  - This illustrates *visual instrumental variables* (VIV)
- GLS applied to equation (12) *is* 2SLS
  - The instruments here are lottery-number indicators. Define $Z_i \equiv \{r_{ji} = 1[\mathrm{R}_i = j]; j = 1, ..., J-1\}$
    - The first stage for D_i on Z_i plus a constant is saturated, so fitted values are cond. means, $\hat{p}_j$, repeated n_j times for each j
    - The second stage slope estimate is therefore weighted least squares on the grouped equation, (12), weighted by n_j
  - Because GLS is efficient, 2SLS is also the efficient linear combination of the underlying just-identified IV (Wald) estimates
- That's why we call Figure 3 "VIV"
  - Sargan/Hansen overid tests the fit of this line
- Fig. 3 also illustrates *two-sample IV*: $\bar{y}_j$ from one smpl, $\hat{p}_j$ from another (details in AK 1992, 1995; Inoue and Solon, 2010)

There's Weakness in Numbers

(of instruments)

2SLS is Biased, Yo

- OLS estimates are unbiased and consistent for the corresponding pop reg (maybe not the reg you want, but nicely estimated)
- 2SLS estimates are *consistent* for causal FX but biased
- Endogenous var. is vector x; dep. var. is vector y; no covs:

$$y = \beta x + \eta \tag{13}$$

The $N \times \mathrm{Q}$ matrix of instruments is Z, with first-stage

$$x = Z\pi + \xi \tag{14}$$

Outcome error η_i is correlated with ξ_i. Instruments are uncorrelated with ξ_i by construction and with η_i by assumption

- The 2SLS estimator is

$$\widehat{\beta}_{2SLS} = \left(x' P_Z x\right)^{-1} x' P_Z y = \beta + \left(x' P_Z x\right)^{-1} x' P_Z \eta$$

where $P_Z = Z(Z'Z)^{-1}Z'$ produces fitted values

Bias and First-stage F

- A Bekker (1994) approximation generates:

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma^2_{\xi}} \frac{1}{F+1} \tag{15}$$

  where

$$F \equiv (1/\sigma^2_{\xi}) E\left(\pi' Z' Z \pi\right) / \mathrm{Q}$$

  is the "population first stage F"
- As F gets small, the bias of 2SLS approaches $\frac{\sigma_{\eta\xi}}{\sigma^2_{\xi}}$
  - The bias of the OLS estimator is $\frac{\sigma_{\eta\xi}}{\sigma^2_{x}}$, which also equals $\frac{\sigma_{\eta\xi}}{\sigma^2_{\xi}}$ if $\pi = 0$
- 2SLS estimates are therefore said to be "biased towards OLS estimates" when the first stage is weak
- The bias of 2SLS vanishes as F increases, as it should when $\pi \neq 0$ and sample size grows

First-stage F (cont.)

- Bias grows as the number of instruments grows (if the instruments are weak)
  - Adding instruments with no effect on the first-stage R-squared, the model sum of squares, $E\left(\pi' Z' Z \pi\right)$, and the residual variance, σ^2_{ξ}, are fixed while Q increases
  - From this we learn that the addition of weak instruments decreases F and therefore increases bias
- Holding the first-stage sum of squares fixed, bias is least in the just-ID case when the number of instruments is as low as it can get
- 2SLS bias is a consequence of first-stage estimation error. We'd like to use $\widehat{x}_{pop} = Z\pi$ as an instrument since these fits are uncorrelated with second stage resids
  - In practice, we use $\widehat{x} = P_Z x = Z\pi + P_Z \xi$
  - 2SLS bias arises from the corr between $P_Z \xi$ and η

IV Without Bias or Tears

- The **reduced form** is unbiased: if the relationship you're after is invisible in the reduced form, then it ain't there!
  - In just-identified models, the p-value for the reduced-form effect of the instrument is approximately the p-value from the second stage
  - Chernozhukov and Hansen (2008) develop reduced-form-based inference for over-identified models
- **LIML** is approximately median-unbiased for constant-effects (but beware heteroskedasticity)
- **Just-identified 2SLS** is approximately unbiased (proof: just-ID=LIML)
  - The just-ID and LIML sampling distributions have no official moments, yet their medians are where they should be
- Split-sample IV (**SSIV**) and jackknife IV **(JIVE)** are Bekker-unbiased (Angrist and Krueger 1995; Angrist, Imbens, and Krueger 1999)
  - Updates include Hausman, Newey, Woutersen, Chao, and Swanson (2012), many others

Monte Carlo for Many-Weak

$$\begin{aligned} y_i &= \beta x_i + \eta_i \\ x_i &= \sum_{j=1}^{\mathrm{Q}} \pi_j z_{ij} + \xi_i \end{aligned}$$

with $\beta = 1$, $\pi_1 = 0.1$, $\pi_j = 0\ \forall j > 1$, joint normal errors with $corr(\eta_i, \xi_i) = .8$, where the instruments, z_{ij}, are independent, standard normals. The sample size is 1000.

- **Figure 4.6.1**: OLS, just identified IV (Q=1, labeled IV; F=11.1), 2SLS (Q=2, labeled 2SLS; F=6.0), LIML (Q=2)
- **Figure 4.6.2**: OLS, 2SLS, and LIML with Q=20 (1 good instrument, 19 worthless; F=1.51)
- **Figure 4.6.3**: OLS, 2SLS, and LIML with Q=20 but $\pi_j = 0$; $j = 1, ..., 20$ (all 20 worthless; F=1.0)
- Quarter of birth estimates of the returns to schooling (reprise): **Table 4.6.2**

Welcome to the Machine

New Models and Methods

- Belloni, Chen, Chernozhukov, and Hansen (2012) use machine learning to pick a few instruments when you're blessed w/an abundance thereof
- The leading ML method here is lasso, a type of "regularized regression", minimizing

$$\min_{\{b_j\}} \underbrace{E[Y_i - \sum_j b_j x_j]^2}_{squared\ \ error} + \underbrace{\frac{\lambda}{n} \sum_j |b_j|}_{penalty\ \ term} \tag{16}$$

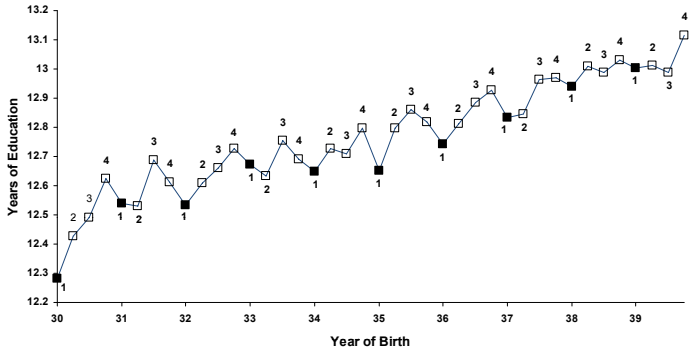
  where λ is a user-chosen penalty
  - Lasso favors lower-dimensional "sparse" models and small coefficients
  - The absolute value inside the penalty term causes lasso to drop some regressors, while shrinking others
- *Post-lasso* runs conventional OLS on the regressors lasso retains
- BCH (2012) discuss the theory behind a post-lasso 2SLS first stage
  - Sounds promising!

What have we found? **The same old fears . . .**

- Sims based on AK91 with 180 instruments (QOB*YOB; QOB*POB) and even 1530 (QOB*YOB*POB) show that **LIML and SSIV beat ML** for bias and MAE
- Lasso for instrument selection faces two challenges
  - 2SLS is (still) biased, yo
  - 2SLS w/a lassoed first stage is pretesting
- Details
  - The good behavior of lasso is predicated on the assumptions of "approximate sparsity," which implies the sample grows relative to the number of first-stage parameters
  - The Bekker sequence reveals the finite sample behavior of 2SLS, SSIV, LIML etc. by fixing the number of obs/parameter; Bekker isn't sparse
  - Hall, et al. (1996) show the dangers of test-based solutions to the weak instruments problem (Andrews, Stock, and Sun 2019 update this)
- Better to use a Bekker-unbiased estimator from the get-go (Angrist and Frandsen 2019)

Tables and Figures

A. Average Education by Quarter of Birth (first stage)



B. Average Weekly Wage by Quarter of Birth (reduced form)

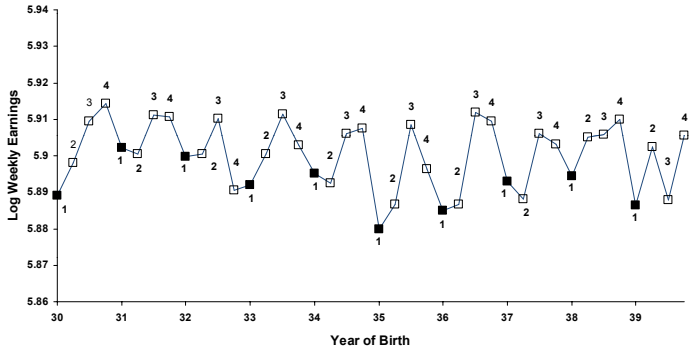


TABLE 4.1.1
2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	.071	.067	.102	.13	.104	.108	.087	.057
	(.0004)	(.0004)	(.024)	(.020)	(.026)	(.020)	(.016)	(.029)
Exogenous Covariates								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year-of-birth dummies		✓			✓	✓	✓	✓
50 state-of-birth dummies		✓			✓	✓	✓	✓
Instruments								
dummy for QOB = 1			✓	✓	✓	✓	✓	✓
dummy for QOB = 2				✓		✓	✓	✓
dummy for QOB = 3				✓		✓	✓	✓
QOB dummies interacted with year-of-birth dummies (30 instruments total)							✓	✓

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the Angrist and Krueger (1991) 1980 census sample. This sample includes native-born men, born 1930–39, with positive earnings and nonallocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QOB denotes quarter of birth.

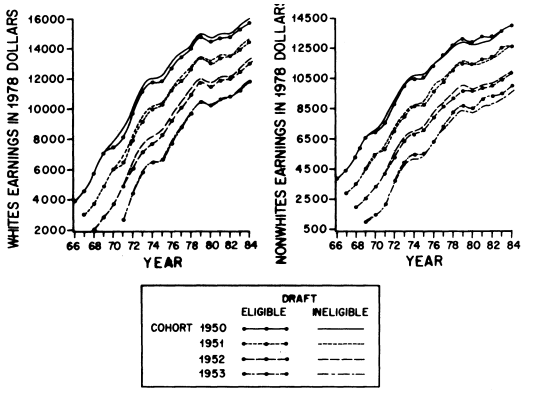


FIGURE 1. SOCIAL SECURITY EARNINGS PROFILES BY DRAFT-ELIGIBILITY STATUS

Notes: The figure plots the history of FICA taxable earnings for the four cohorts born 1950–3. For each cohort, separate lines are drawn for draft-eligible and draft-ineligible men. Plotted points show average real (1978) earnings of working men born in 1953, real earnings +$3000 for men born in 1950, real earnings +$2000 for men born in 1951, and real earnings +$1000 for men born in 1952.

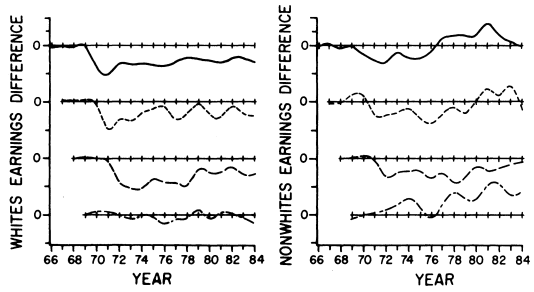


Table 4.1.3

IV Estimates of the Effects of Military Service on the Earnings of White Men born in 1950

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Inelig. Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8 (210.5)	.182	.159 (.040)	-2,741 (1,324)
1971	3,338	-325.9 (46.6)			-2050 (293)
1969	2,299	-2.0 (34.5)			

Note: Adapted from Table 5 in Angrist and Krueger (1999) and author tabulations. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Vete status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

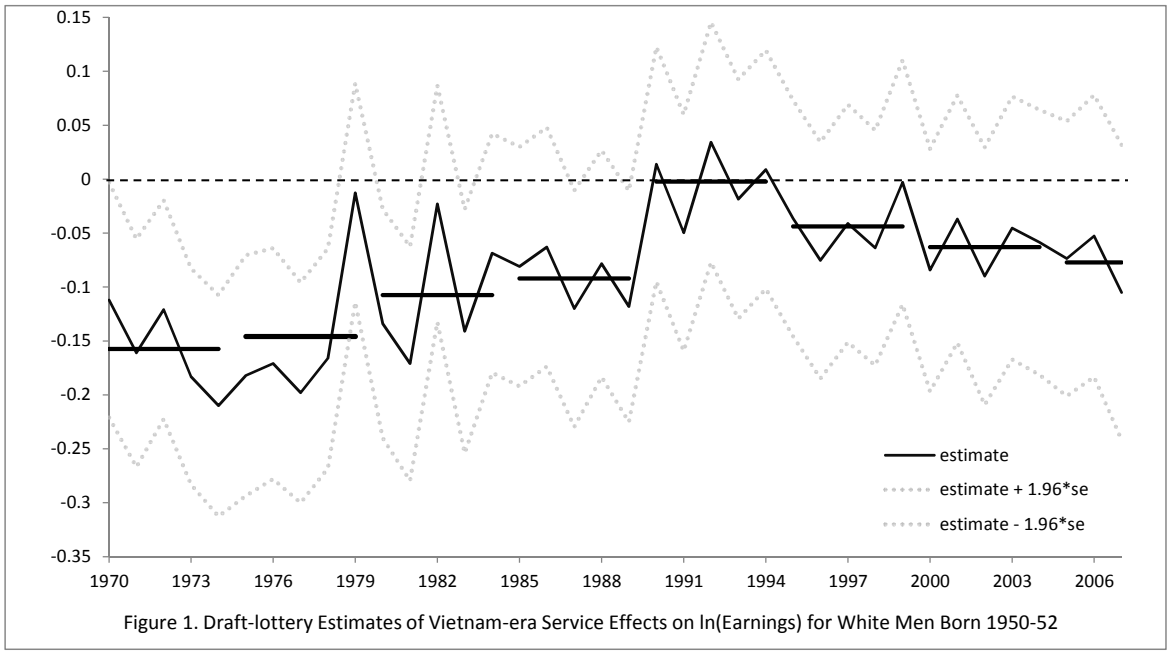


Figure 1. Draft-lottery Estimates of Vietnam-era Service Effects on ln(Earnings) for White Men Born 1950-52

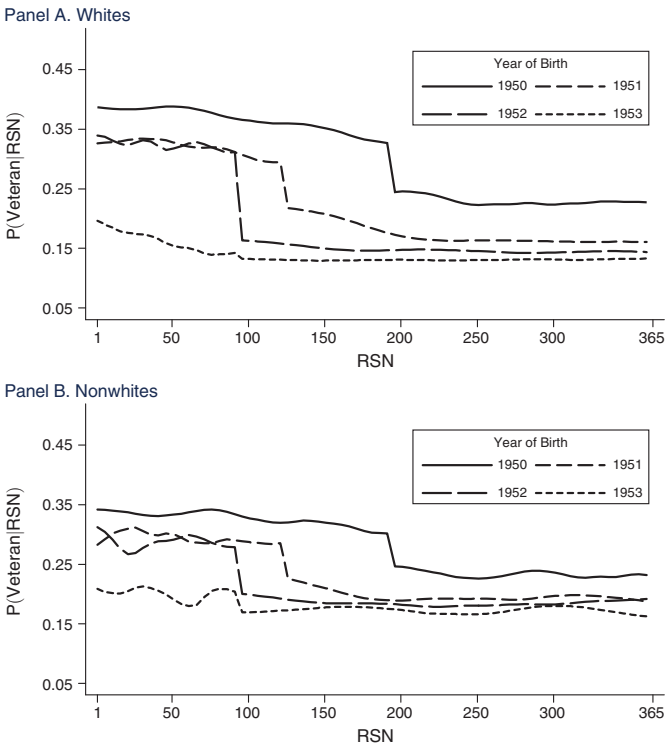


Figure 1. The Conditional Probability of Military Services by Random Sequence Number

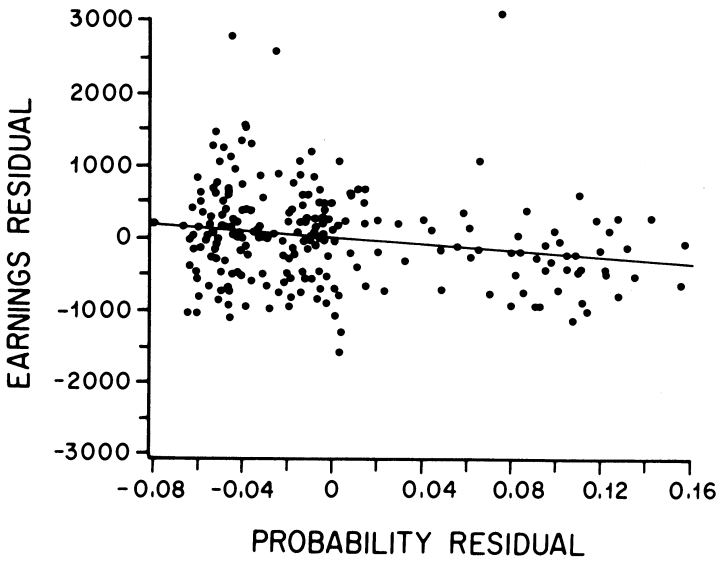


FIGURE 3. EARNINGS AND THE PROBABILITY OF VETERAN STATUS BY LOTTERY NUMBER

Notes: The figure plots mean W-2 compensation in 1981–4 against probabilities of veteran status by cohort and groups of five consecutive lottery numbers for white men born 1950–3. Plotted points consist of the average residuals (over four years of earnings) from regressions on period and cohort effects. The slope of the least-squares regression line drawn through the points is −2,384, with a standard error of 778, and is an estimate of α in the equation

$$\bar{y}_{ctj} = \beta_c + \delta_t + \hat{p}_{cj}\alpha + \bar{u}_{ctj}.$$

4.7. APPENDIX

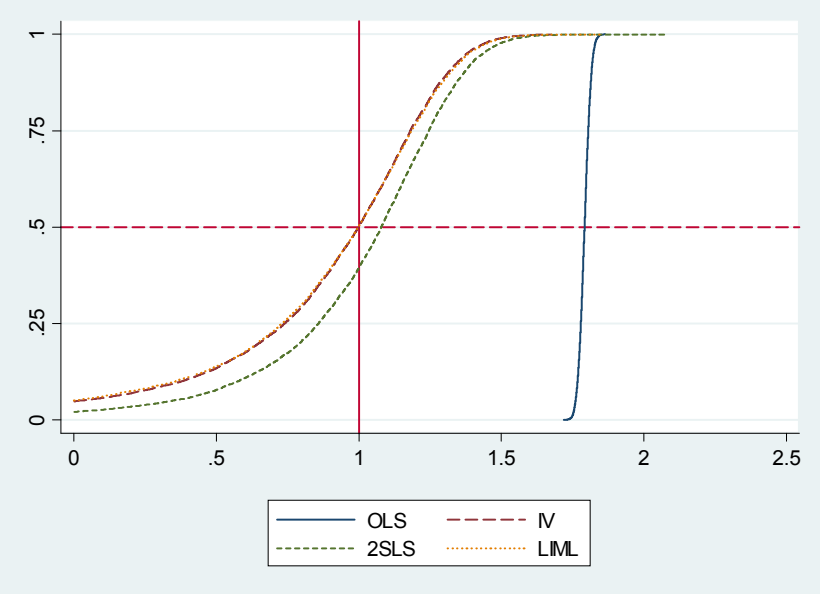


Figure 4.6.1: Distribution of the OLS, IV, 2SLS, and LIML estimators. IV uses one instrument, whil and LIML use two instruments.

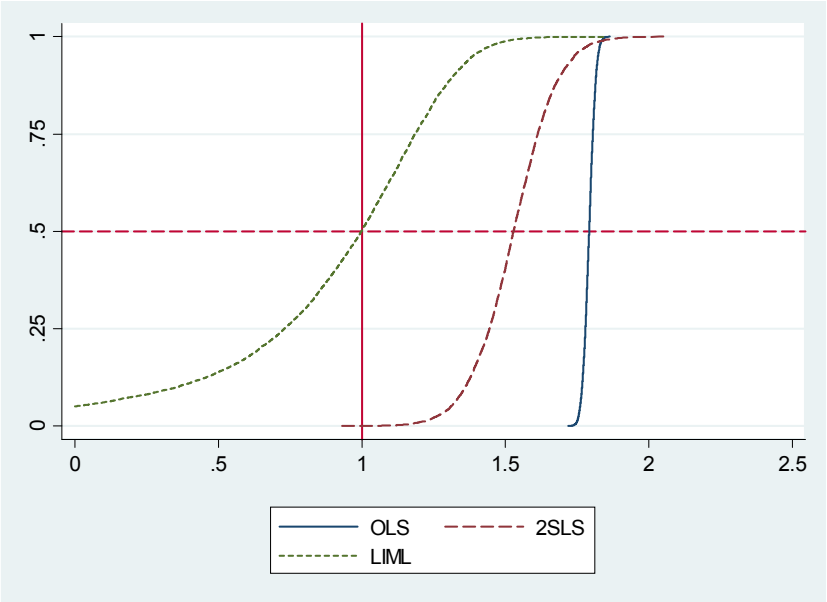


Figure 4.6.2: Distribution of the OLS, 2SLS, and LIML estimators with 20 instruments

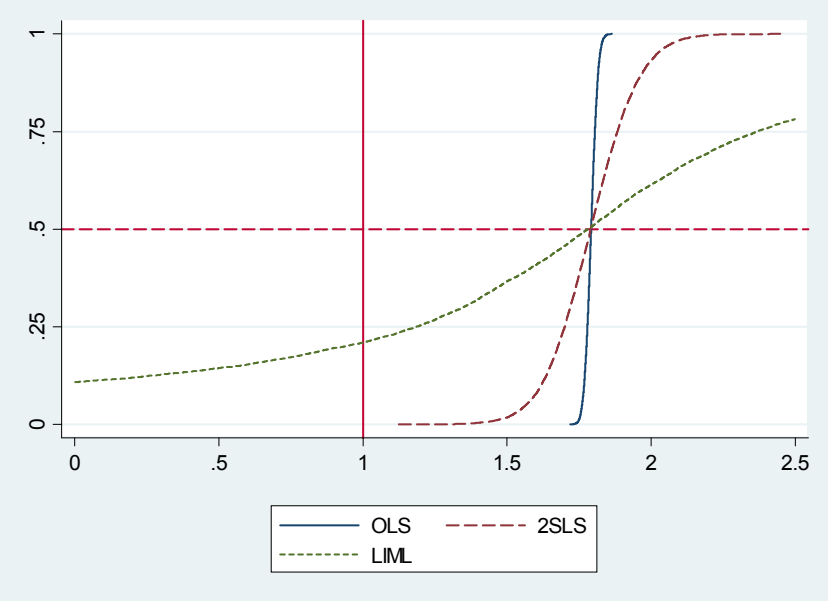


Figure 4.6.3: Distribution of the OLS, 2SLS, and LIML estimators with 20 worthless instruments

TABLE 4.6.2
Alternative IV estimates of the economic returns to schooling

	(1)	(2)	(3)	(4)	(5)	(6)
2SLS	.105	.435	.089	.076	.093	.091
	(.020)	(.450)	(.016)	(.029)	(.009)	(.011)
LIML	.106	.539	.093	.081	.106	.110
	(.020)	(.627)	(.018)	(.041)	(.012)	(.015)
F-statistic (excluded instruments)	32.27	.42	4.91	1.61	2.58	1.97
Controls						
Year of birth	✓	✓	✓	✓	✓	✓
State of birth					✓	✓
Age, age squared		✓		✓		✓
Excluded instruments						
Quarter-of-birth dummies	✓	✓				
Quarter of birth*year of birth			✓	✓	✓	✓
Quarter of birth*state of birth					✓	✓
Number of excluded instruments	3	2	30	28	180	178

Notes: The table compares 2SLS and LIML estimates using alternative sets of instruments and controls. The age and age squared variables measure age in quarters. The OLS estimate corresponding to the models reported in columns 1–4 is .071; the OLS estimate corresponding to the models reported in columns 5 and 6 is .067. Data are from the Angrist and Krueger (1991) 1980 census sample. The sample size is 329,509. Standard errors are reported in parentheses.

	180 Instruments (QOB*YOB; POB*YOB; Average F=2.5)					1530 Instruments (QOB*YOB*POB; Average F=1.7)				
Estimator	Avg. IVs retained (1)	Bias (2)	Standard deviation (3)	Median abs. dev. (4)	Median abs. error (5)	Avg. IVs retained (6)	Bias (7)	Standard deviation (8)	Median abs. dev. (9)	Median abs. error (10)
OLS		0.107	0.0004	0.0003	0.1070					
2SLS	180	0.0403	0.0108	0.0075	0.0397	1530	0.0611	0.0046	0.0032	0.0611
Post-lasso IV (CV penalty)	74.0	0.0390	0.0120	0.0082	0.0384	99.0	0.0559	0.0084	0.0059	0.0560
Post-lasso IV (plug-in penalty, IVs selected)*	2.1	0.0143	0.0346	0.0218	0.0279	1.6	0.0149	0.0367	0.0224	0.0271
Split-Sample IV	180	-0.0009	0.0237	0.0158	0.0158	1530	-0.0001	0.0164	0.0112	0.0115
Post-lasso SSIV (CV penalty)	63.1	-0.0015	0.0258	0.0172	0.0173	63.0	-0.0013	0.0280	0.0183	0.0183
Post-lasso SSIV (plug-in penalty, IVs selected)**	2.1	-0.0724	1.3168	0.0274	0.0287	3.4	0.0197	0.0504	0.0228	0.0292
Post-lasso (IV choice split only, CV penalty)	63.1	0.0429	0.0144	0.0097	0.0431	63.0	0.0460	0.0141	0.0093	0.0459
LIML	180	-0.0016	0.0185	0.0123	0.0124	1530	-0.0034	0.0117	0.0079	0.0083
Post-lasso LIML (CV penalty)	74.0	0.0222	0.0152	0.0102	0.0220	99.0	0.0484	0.0094	0.0066	0.0483
Post-lasso LIML (plug-in penalty, IVs selected)*	2.1	0.0126	0.0347	0.0221	0.0273	1.6	0.0138	0.0366	0.0221	0.0257
Pretested LIML (t => 3.12 for 180, t=>2.3 for 1530)	18	0.0222	0.0236	0.0148	0.0238	153	0.0385	0.0163	0.0111	0.0393
Random forest first stage, 2SLS using RF fits as instruments (min leaf size=1)							0.0611	0.0047	0.0030	0.0612
Random forest 2SLS, min leaf size = 800							0.0567	0.0065	0.0045	0.0567
Random forest first stage, SSIV using RF fits as instruments (min leaf size =1)							-0.0003	0.0158	0.0109	0.0108
Random forest SSIV, min leaf size = 800							-0.0005	0.0158	0.0104	0.0103

Notes: The table describes simulation results for 999 Monte Carlo estimates of the economic returns to schooling using simulated samples constructed from the Angrist and Krueger (1991) census sample of men born 1930-39 (N=329,509). The causal effect of schooling is calibrated to 0.1; the OLS estimand is 0.207. The instruments used to compute the estimates described by columns 1-5 consist of 30 quarter-of-birth-by-year-of-birth and 150 quarter-of-birth-by-state-of-birth interactions (average F-stat = 2.5, average concentration parameter = 270). The instruments used to compute the estimates described by columns 6-10 are quarter-of-birth-by-year-of-birth-by-state-of-birth interactions (average F-stat = 1.7, average concentration parameter = 1050). All models include saturated year of birth by state of birth controls. Columns 1 and 6 report the average number of instruments retained by lasso. Post-lasso estimates are computed as described in the appendix. Split-Sample IV uses first stage coefficients estimated in one half-sample to construct a cross-sample fitted value used for IV in the other. Sample-splitting procedures average results from complementary splits. Post-lasso with an IV-choice split only uses post-lasso in half the sample to pick instruments, doing 2SLS with these and own-sample fitted values in the other half. "Post-lasso LIML" is LIML using the instrument set selected by a post-lasso first stage. "Pretested LIML" estimates are computed using conventional LIML, retaining only instruments with a first-stage t-statistic in the upper decile of t-statistics for the full set of instruments. Simulation sets choose lasso penalties once, using the original AK91 data. Random forest routines are described in the appendix.
*The plug-in penalty generates a lasso first stage that includes no instruments in 11 simulation runs with 180 instruments and in 57 simulation runs with 1530