INSTRUMENTAL VARIABLES (Take 1): CONSTANT EFFECTS

Josh Angrist

MIT 14.387 (Fall 2014)

Organizing IV

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

- The constant effects framework focuses attention on the IV solution for selection bias and on essential IV mechanics
- But first: Why do IV?
  - I can't say "because the regressors are correlated with the errors."
  - As we've seen, regressors are uncorrelated with errors by definition
- The (short) regression of schooling on wages produces residuals uncorrelated with schooling (that's how the good lord made 'em)
- The problem, therefore, must be that the regression you've got is not the regression you want (and that's your fault!)

IV Goes Long

- Suppose the causal link between schooling and wages can be written $f_i(s) = \alpha + \rho s + \eta_i$
- Imagine a vector of control variables, A_i, called "ability"; write

$$\eta_i = A_i'\gamma + v_i$$

where γ is a vector of pop. reg. coefficients, so v_i and A_i are uncorrelated *by construction*

- We'd happily include ability in the regression of wages on schooling, producing this long regression:

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

The error term here is the random part of potential outcomes, v_i, left over after controlling for A_i

- If $E[\mathrm{S}_i v_i] = 0$, a version of the CIA, the population regression of Y_i on S_i and A_i identifies ρ. That's like saying: "A_i is the *only* reason schooling is correlated with potential outcomes."

IV and OVB

- *IV allows us to estimate the long-regression coefficient, ρ, when A_i is unobserved.*

  The instrument, Z_i, is assumed to be: (1) correlated with the causal variable of interest, S_i; and (2) uncorrelated with potential outcomes
- Here, "uncorrelated with potentials" means $Cov(\eta_i, \mathrm{Z}_i) = 0$, or, equivalently, Z_i is uncorrelated with both A_i and v_i
  - This is a version of the *exclusion restriction*: Z_i can be said to be excluded from the causal model of interest
- Given the exclusion restriction, it follows from equation (1) that

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

- The IV *estimator* is the sample analog of (2)

Two-stage least squares (2SLS)

- In practice, we do IV by doing 2SLS. This allows us to add covariates (controls) and combine multiple instruments. Returning to the schooling example, a causal model with covariates is

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho \mathrm{S}_i + \eta_i, \tag{3}$$

where η_i is the compound error term, $A_i\gamma + v_i$. The first stage and reduced form are

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_i' \pi_{10} + \pi_{11} \mathrm{Z}_i + \xi_{1i} && (4) \\ \mathrm{Y}_i &= \mathrm{X}_i' \pi_{20} + \pi_{21} \mathrm{Z}_i + \xi_{2i} && (5) \end{aligned}$$

- The reduced form is obtained by substituting (4) into (3):

$$\begin{aligned} \mathrm{Y}_i &= \alpha' \mathrm{X}_i + \rho[\mathrm{X}_i' \pi_{10} + \pi_{11} \mathrm{Z}_i] + \rho \xi_{1i} + \eta_i && (6) \\ &= \mathrm{X}_i' [\alpha + \rho \pi_{10}] + \rho \pi_{11} \mathrm{Z}_i + [\rho \xi_{1i} + \eta_i] \\ &= \mathrm{X}_i' \pi_{20} + \pi_{21} \mathrm{Z}_i + \xi_{2i} \end{aligned}$$

2SLS Notes

- Again, it's all about the ratio of RF to 1st:

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

  In simultaneous squations models, the sample analog of this ratio is called an *Indirect Least Squares* (ILS) estimator of ρ

- Where does *two-stage least squares* come from? Write the first stage as the sum of fitted values plus first-stage residuals:

$$\mathrm{S}_i = \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i + \xi_{1i} = \hat{s}_i + \xi_{1i}$$

2SLS estimates of (3) can be constructed by substituting first-stage fitted values for S_i in (3):

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\hat{s}_i + [\eta_i + \rho\xi_{1i}], \tag{7}$$

and using OLS to estimate this "second stage" (a version of eq. 6)

- In practice, we let Stata do it: "manual 2SLS" doesn't get the standard errors right

2SLS example: Angrist and Krueger (1991)

- AK-91 argue that because children born in late-quarters start school younger, they are kept in school longer by birthday-based compulsory schooling laws
- There's a powerful first stage supporting this: Schooling tends to be higher for late-quarter births; this is driven by high school and not college, consistent with the CSL story
- The QOB first stage and reduced form are plotted in **Figure 4.1.1**
- The corresponding 2SLS estimates appear in **Table 4.1.1**
  - 2SLS matches the QOB pattern earnings (the RF) to the QOB pattern in schooling (the first stage).
  - The exogenous covariates include year-of-birth and state-of-birth dummies, as well as linear and quadratic functions of age in quarters
- QOB Questioned: Bound, Jaeger, and Baker (1995) and Buckles and Hungerman (2008) argued QOB is correlated with maternal characteristics. **Allowing for this** fails to overturn AK conclusions

2SLS is a many-splendored thing

- 2SLS is the same as IV where the instrument is $\hat{s}_i^*$, the residual from a regression of $\hat{s}_i$ on X_i
- One-instrument 2SLS equals IV, where the instrument is $\tilde{z}_i$, the residual from a regression of Z_i on the covs, X_i
- One-instrument 2SLS equals indirect least squares (ILS), that is, the ratio of reduced form to first stage coefficients on the instrument. In other words,

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

- With more than one instrument, 2SLS is a weighted average of the one-at-time (just-identified) estimates (In a linear homoskedastic constant-effects model, this is efficient)

Multi-Instrument 2SLS (details; **mistakes**)

- Let

$$\rho_j = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_{ji})}{Cov(\mathrm{D}_i, \mathrm{Z}_{ji})}; j = 1, 2$$

  denote two IV estimands using Z_{1i} and Z_{2i} to instrument D_i.

- The 2SLS estimand is

$$\rho_{2SLS} = \psi\rho_1 + (1 - \psi)\rho_2,$$

  where ψ is a number between zero and one that depends on the relative strength of the instruments in the first stage.

- Angrist and Evans (1998) use twins and sex-mix instruments
  - Using a twins-2 instrument alone, the IV estimate of the effect of a third child on female labor force participation is -.084 (s.e.=.017). The corresponding samesex estimate is -.138 (s.e.=.029).
  - Using both instruments produces a 2SLS estimate of -.098 (.015).
  - The 2SLS weight in this case is .74 for twins, .26 for samesex, due to the stronger twins first stage.

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 have yet to be relegated to the dustbin of IV history:
  - Manual 2SLS
  - Covariate ambivalence
  - Forbidden regressions (from the left and the right)
- These can be interpreted as the result of failed attempts to get round hard-wired 2SLS protocols
- Avoid temptation: let Stata do it!

2SLS Lingo

- These terms come to us from simultaneous equations modeling, the intellectual birthplace of IV:
  - *Endogenous variables* are the dependent variable and the independent variable(s) to be instrumented; in a simultaneous equations model, endogenous variables are determined by solving the system
  - To *treat an independent variable as endogenous* is to instrument it, i.e., to replace it with fitted values in the 2SLS second stage
  - *Exogenous variables* include *covariates* (not instrumented) and the excluded instruments themselves. In a simultaneous equations model, exogenous variables are determined outside the system
- In any IV study, variables are either: dependent or (other) endogenous variables, instruments, or covariates
- If you're unsure what's what, or find yourself asking variables to play more than one role . . . seek counseling

The Wald estimator

- How were Vietnam-era vets affected by their service?
- Let D_i indicate veterans. A causal constant-effects model is:

$$\mathrm{Y}_i = \alpha + \rho \mathrm{D}_i + \eta_i, \tag{8}$$

where η_i and D_i may be correlated. B/c Z_i is a dummy,

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

with an analogous formula for $\frac{Cov(\mathrm{D}_i, \mathrm{Z}_i)}{V(\mathrm{Z}_i)}$. It follows that,

$$\rho = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{Cov(\mathrm{D}_i, \mathrm{Z}_i)} = \frac{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]}{E[\mathrm{D}_i|\mathrm{Z}_i = 1] - E[\mathrm{D}_i|\mathrm{Z}_i = 0]} \tag{9}$$

- A direct route uses (8) and $E[\eta_i|\mathrm{Z}_i] = 0$:

$$E[\mathrm{Y}_i|\mathrm{Z}_i] = \alpha + \rho E[\mathrm{D}_i|\mathrm{Z}_i] \tag{10}$$

Solving this for ρ produces (9)

Earnings Consequences of Vietnam-Era Military Service (Angrist, 1990)

- Key variables

$$\begin{aligned} Z_i &= \text{randomly assigned draft-eligibility in the 1970-72 draft lotteries} \\ D_i &= \text{a dummy indicating Vietnam-era veterans} \end{aligned}$$

- The causal effect of Vietnam-era military service is the difference in average earnings by draft-eligibility status (RF) divided by the difference in the probability of service (first stage):

$$\begin{aligned} \frac{Cov(D_i, Z_i)}{V(Z_i)} &= E[D_i|Z_i = 1] - E[D_i|Z_i = 0] \\ &= P[D_i = 1|Z_i = 1] - P[D_i = 1|Z_i = 0] \end{aligned}$$

- See RF, first stage, and IV in **Angrist (1990), Figures 1-2 and MHE Table 4.1.3** (based on Angrist 1990,Table 3).
- Draft lottery updates: **Angrist, Chen, and Song (2011)**

Multiple groups and 2SLS

- There's more to the draft lottery than draft-eligibility: **Angrist and Chen (2008), Figure 1**
- Let R_i denote draft lottery numbers. The draft-eligibility Wald estimator uses $1[\mathrm{R}_i < 195]$ as an instrument in a just-identified setup
- The first stage that uses everything we know can be written:

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

since $P[\mathrm{D}_i = 1|\mathrm{R}_i] = E[\mathrm{D}_i|\mathrm{R}_i]$. Suppose $\mathrm{R}_i \in j = 1, ..., \mathrm{J}$. We can estimate ρ using J grouped obs by fitting

$$\bar{y}_j = \alpha + \rho \hat{p}_j + \bar{\eta}_j \tag{12}$$

- Efficient GLS for grouped data in a constant-effects linear model is weighted least squares, in this case weighted by the variance of $\bar{\eta}_j$ (Prais and Aitchison, 1954). If η_i has variance σ^2_η, the grouped variance is $\frac{\sigma^2_\eta}{n_j}$, where n_j is the group size.

Visual IV, Grouping, and GLS

- Equation (12) in action: **Angrist (1990), Figure 3**. This illustrates visual instrumental variables (VIV)
- GLS (weighted least squares) applied to equation (12) *is* 2SLS
  - The instruments in this case are dummies for each lottery-number cell. 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 the cell size, n_j
  - Because GLS is efficient, 2SLS is also the efficient linear combination of the underlying just-identified IV (Wald) estimates (earlier, we saw that 2SLS is a weighted average of just-identified estimates in a two-instrument example)
- That's why we call Figure 3 "VIV"

Specification Testing [**TSIV**]

Suppose the residuals, η_i, are conditionally homoskedastic. GLS on the grouped equation, (12), chooses parameter estimates a and b to minimize

$$\hat{J}_N(a,b) = (1/\sigma_\eta^2) \times \sum_j n_j(\bar{y}_j - a - b\hat{p}_j)^2 \tag{13}$$

(If the residuals are heteroskedastic, replace constant σ_η^2 with σ_j^2, the variance of η_i in group j)

- The minimized GLS minimand is the (Sargan) over-id test statistic for 2SLS estimates constructed using group dummies as instruments (MHE 4.2.2). This test statistic is distributed $\chi^2\,(\mathrm{J}-2)$ if J groups are used to estimate a slope and intercept
- Over-id for dummy instruments measures the fit of the line connecting $\bar{y}_j$ and $\hat{p}_j$ in a VIV plot like Fig. 3 in Angrist (1990)
- The over-identification test statistic is also a (Wald) test statistic for equality of a full set of linearly independent Wald estimates (implied by Newey and West (1987); see Angrist (1991))

Two-Sample IV

- Let $\{\mathrm{Y}_j, W_j, Z_j; j = 1, 2\}$ be data from two samples, where W_j and Z_j include exog covs. Angrist (1990) constructs (first-stage) $\frac{Z_2' W_2}{N_2}$ from military records, while Social Security records were used to construct (reduced form) $\frac{Z_1' \mathrm{Y}_1}{N_1}$
- AK-95 and Inoue and Solon (2010) simplify: First-stage fits in ds2 are $(Z_2' Z_2)^{-1} Z_2' W_2$. Carry over to ds1 by constructing the *cross-sample fitted value*, $\hat{W}_{12} \equiv Z_1 (Z_2' Z_2)^{-1} Z_2' W_2$. The second stage for this version of TSIV (which AK-95 call SSIV) regresses Y_1 on $\hat{W}_{12}$. The cross-sample fitted value is

$$\hat{w}_{12,i} = \mathrm{z}_{1i} \hat{\pi}_2,$$

  where $\hat{\pi}_2$ is the first-stage effect estimated using ds2 and the z_{1i} $(i = 1, ..., N_1)$ are the instruments in ds1
- Manual 2SLS, yikes! Inoue and Solon (2010) get the standard errors right, among other TSIV improvements

The Bias of 2SLS

- Cross-section OLS estimates are typically unbiased for the pop BLP, as well as consistent, but this might not be the regression you want
- 2SLS estimates are consistent for causal effects but biased towards OLS estimates
- Endogenous var. is vector x; dep. var. is vector y; no covs:

$$y = \beta x + \eta \tag{14}$$

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

$$x = Z\pi + \xi \tag{15}$$

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 $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix that produces fitted values

The Bias of 2SLS (cont.)

- Substituting for x in $x'P_Z\eta$, we get

$$\widehat{\beta}_{2SLS} - \beta = \left(x'P_Z x\right)^{-1}\left(\pi'Z' + \xi'\right)P_Z\eta \quad (16)$$

$$= \left(x'P_Z x\right)^{-1}\pi'Z'\eta + \left(x'P_Z x\right)^{-1}\xi'P_Z\eta \quad (17)$$

- Expectation of the ratios on the right hand side of (16) are closely approximated by the ratio of expectations:

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \left(E[x'P_Z x]\right)^{-1}E[\pi'Z'\eta] + \left(E[x'P_Z x]\right)^{-1}E[\xi'P_Z\eta].$$

  This Bekker (1994) approximation ("group asymptotics" in AK-95) gives a good account of finite-sample behavior
- Using the fact that $E[\pi'Z'\xi] = 0$ and $E[\pi'Z'\eta] = 0$, we have

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \left[E\left(\pi'Z'Z\pi\right) + E(\xi'P_Z\xi)\right]^{-1}E\left(\xi'P_Z\eta\right) \quad (18)$$

- 2SLS is biased b/c $E\left(\xi'P_Z\eta\right) \neq 0$ unless η_i and ξ_i are uncorrelated

The Bias of 2SLS: First-stage F

- Manipulation of (18) generates:

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma_{\xi}^2} \left[\frac{E\left(\pi' Z' Z \pi\right)/Q}{\sigma_{\xi}^2} + 1 \right]^{-1}$$

$(1/\sigma_{\xi}^2) E\left(\pi' Z' Z \pi\right)/Q$ is the "population F" for joint significance of instruments in first-stage, so we can write

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma_{\xi}^2} \frac{1}{F+1} \tag{19}$$

- As F gets small, the bias of 2SLS approaches $\frac{\sigma_{\eta\xi}}{\sigma_{\xi}^2}$. The bias of the OLS estimator is $\frac{\sigma_{\eta\xi}}{\sigma_x^2}$, which also equals $\frac{\sigma_{\eta\xi}}{\sigma_{\xi}^2}$ if $\pi = 0$. 2SLS estimates are therefore said to be "biased towards OLS estimates" when there isn't much of a first stage. On the other hand, the bias of 2SLS vanishes when F gets large, as it should happen in large samples when $\pi \neq 0$.

The Bias of 2SLS: First-stage F (cont.)

- First-stage F varies inversely with the number of instruments if they're 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
  - The F-statistic shrinks as a result. From this we learn that the addition of weak instruments 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 IVs since these fits are uncorrelated with the second stage error
  - In practice, we use $\widehat{x} = P_Z x = Z\pi + P_Z \xi$, which differs from $\widehat{x}_{pop}$ by the term $P_Z \xi$
  - 2SLS bias arises from the corr between $P_Z \xi$ and η

IV without bias or tears

- **Just-identified 2SLS** (say, the Wald estimator) is *approximately unbiased* (this isn't clear from the Bekker sequence). The just-ID sampling distribution has no moments, yet it's approximately centered where it should be unless the instruments are *really* weak
- The **Reduced Form** is unbiased: if you can't see the relationship you're after in the reduced form, 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, use this to do bias-free inference)
- **LIML** is approximately median-unbiased for over-identified constant-effects models, and therefore provides an attractive alternative to just-identified estimation using one instrument at a time **(**see, e.g., Davidson and MacKinnon, 1993, and Mariano, 2001**).** **(**LIML=2SLS in just-identified models**)**

Alternative Estimators

- Riff on the SSIV idea: JIVE (Angrist, Imbens, and Krueger, 1999) removes bias using leave-out first stage fits for each observation (the fitted value for observation i is $Z_i\hat{\pi}_{(i)}$ where $\hat{\pi}_{(i)}$ is an estimate that throws out observation i). JIVE sounds appealing, but AIK and others have found that it rarely beats LIML)
- The right linear combination of OLS and 2SLS should be approximately unbiased. It turns out that LIML is just such a "combination estimator" (see the working paper version of AIK-99). You might also try Fuller's (1977) modified LIML, discussed by Hahn, Hausman, and Kuersteiner (2004). Fuller may be more precise than LIML in finite samples since it has moments
- Hausman, *et al. (2007)* modify LIML and Fuller to deal with heteroscedasticity
- Kolesar *et al.* (2011) modify LIML to allow for random effects

Monte Carlo

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

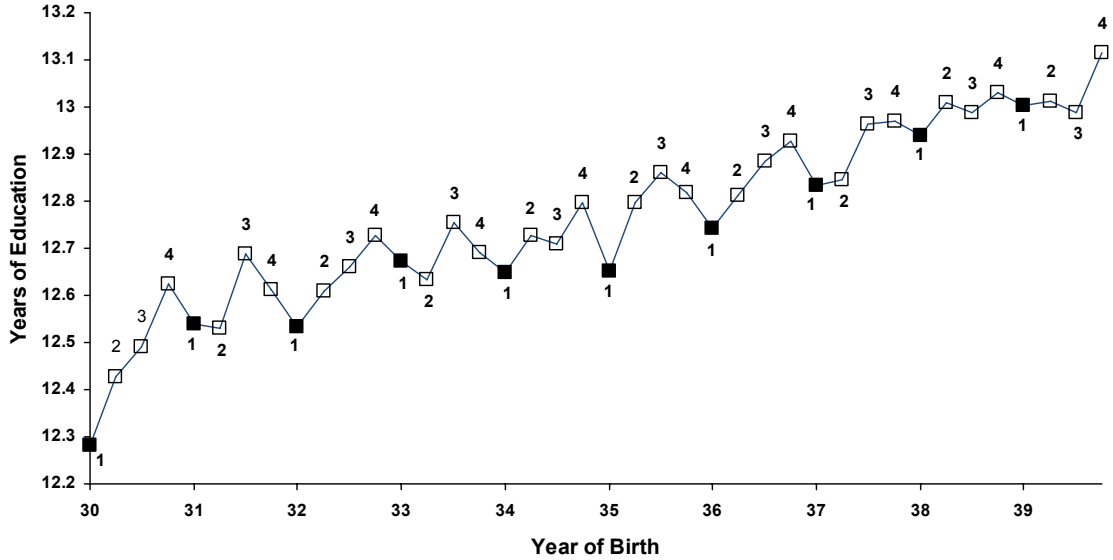
with $\beta = 1$, $\pi_1 = 0.1$, $\pi_j = 0\ \forall j > 1$, joint normal errors with $corr(\eta_i, \xi_i) = .8$, where the instruments, z_{ij}, are independent, standard normals. The sample size is 1000.

- **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**

Tables and Figures

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



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

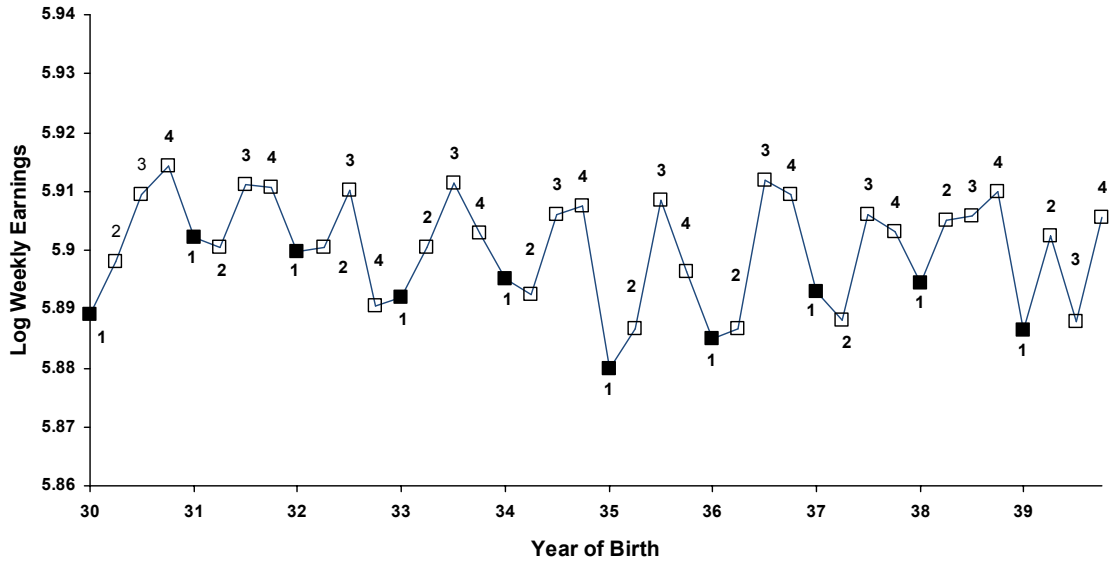


TABLE 4.1.1
2SLS estimates of the economic returns to schooling

	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.

TABLE 6. IV REGRESSIONS ON RETURNS TO EDUCATION: RESULTS FROM THE CENSUS

	Wages, Logged: QOB Instruments		**Wages, Logged: Year*QOB Instruments**		**Wages, in Levels: QOB Instruments**		**Wages, in Levels: Year*QOB Instruments**	
Years of Education	0.103	0.147	0.075	0.09	33.16	49.16	24.13	30.94
	[0.083]	[0.081]	[0.040]	[0.040]	[24.21]	[23.5]	[11.55]	[10.59]
Family Controls?	No	Yes	No	Yes	No	Yes	No	Yes
Instruments	QOB	QOB	YOB*QOB	YOB*QOB	QOB	QOB	YOB*QOB	YOB*QOB
Age Controls?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Dummies?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummies?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weights?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors in brackets. Observations are county-of-birth/quarter-of-birth/year-of-birth cells and all regressions weight by total individuals reporting positive earnings in a cell. The dependent variable in the first two pairs of regressions is the log of average wages in a cell, in the last two pairs of regressions it is the average of cell wages in levels. Regressions are from cohorts of males born between 1944 and 1960; see Table 5 for a description of family characteristic and wage and age variables.

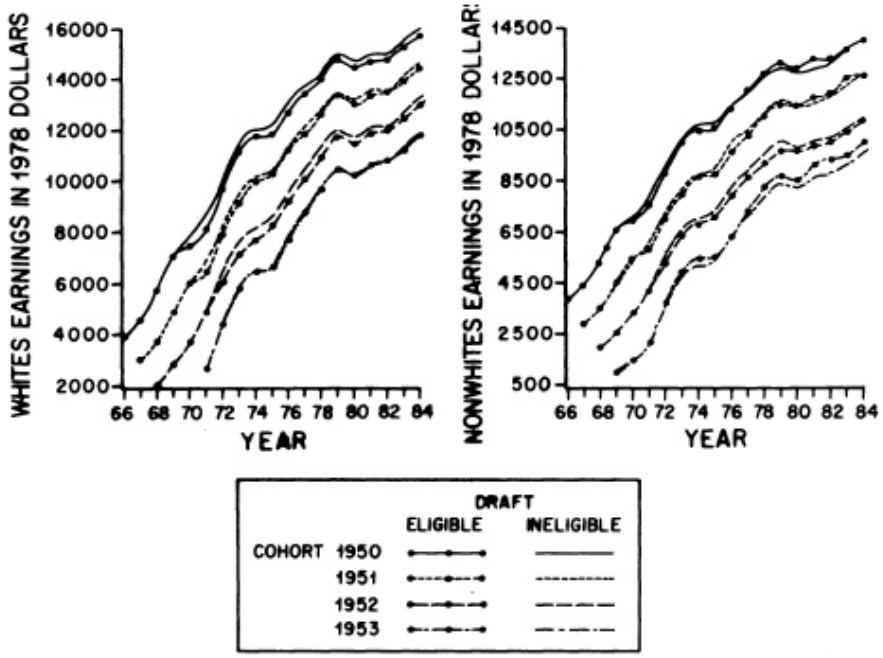


FIGURE 1. SOCIAL SECURITY EARNINGS PROFILES BY DRAFT-ELIGIBILITY STATUS

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.

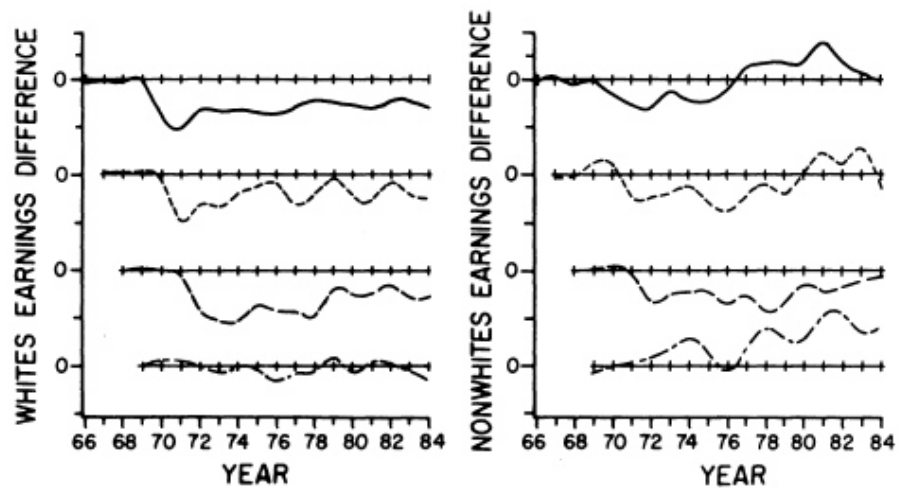


Table 4.1.3

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	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8 (210.5)	.267	.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.

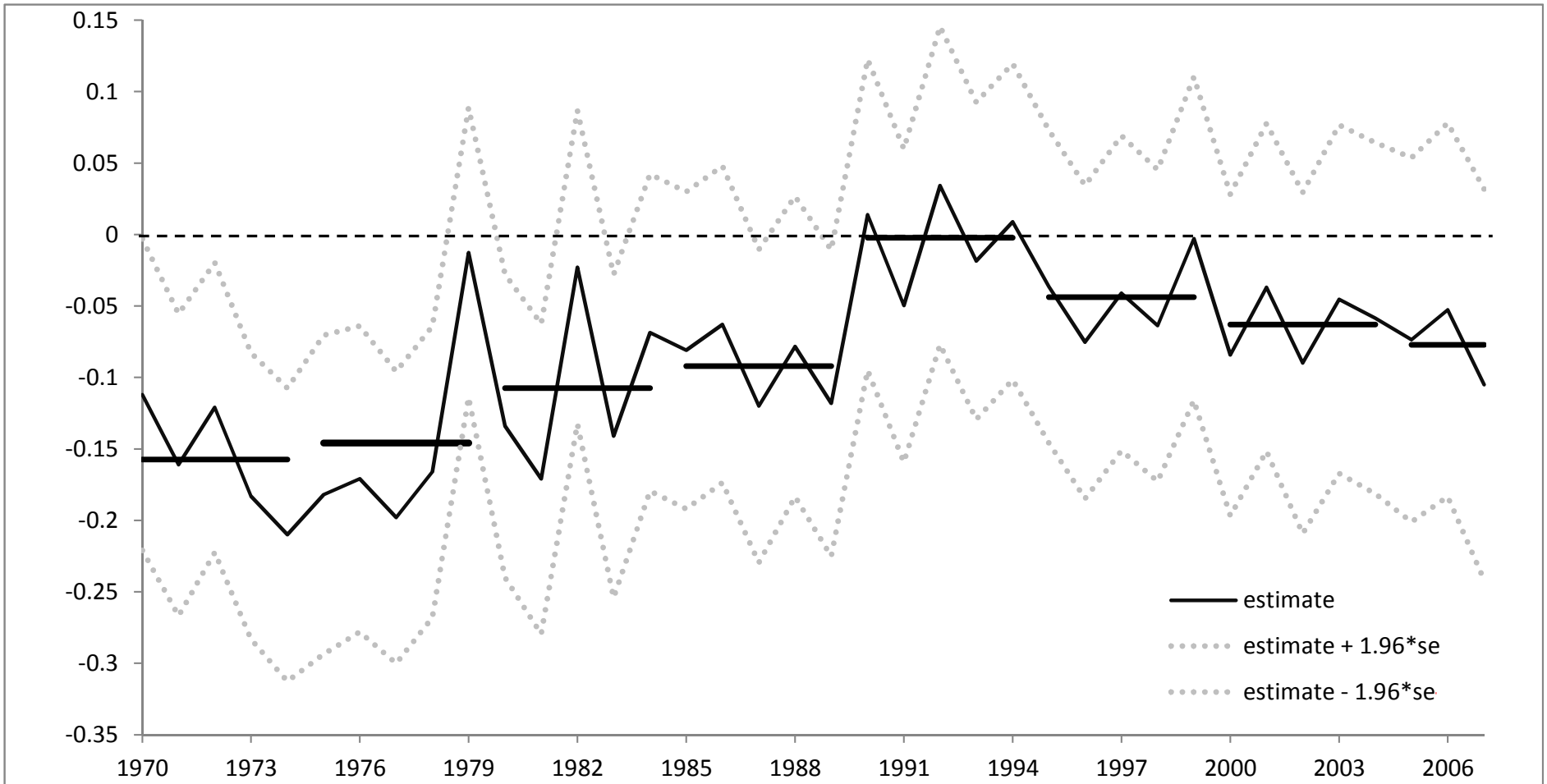
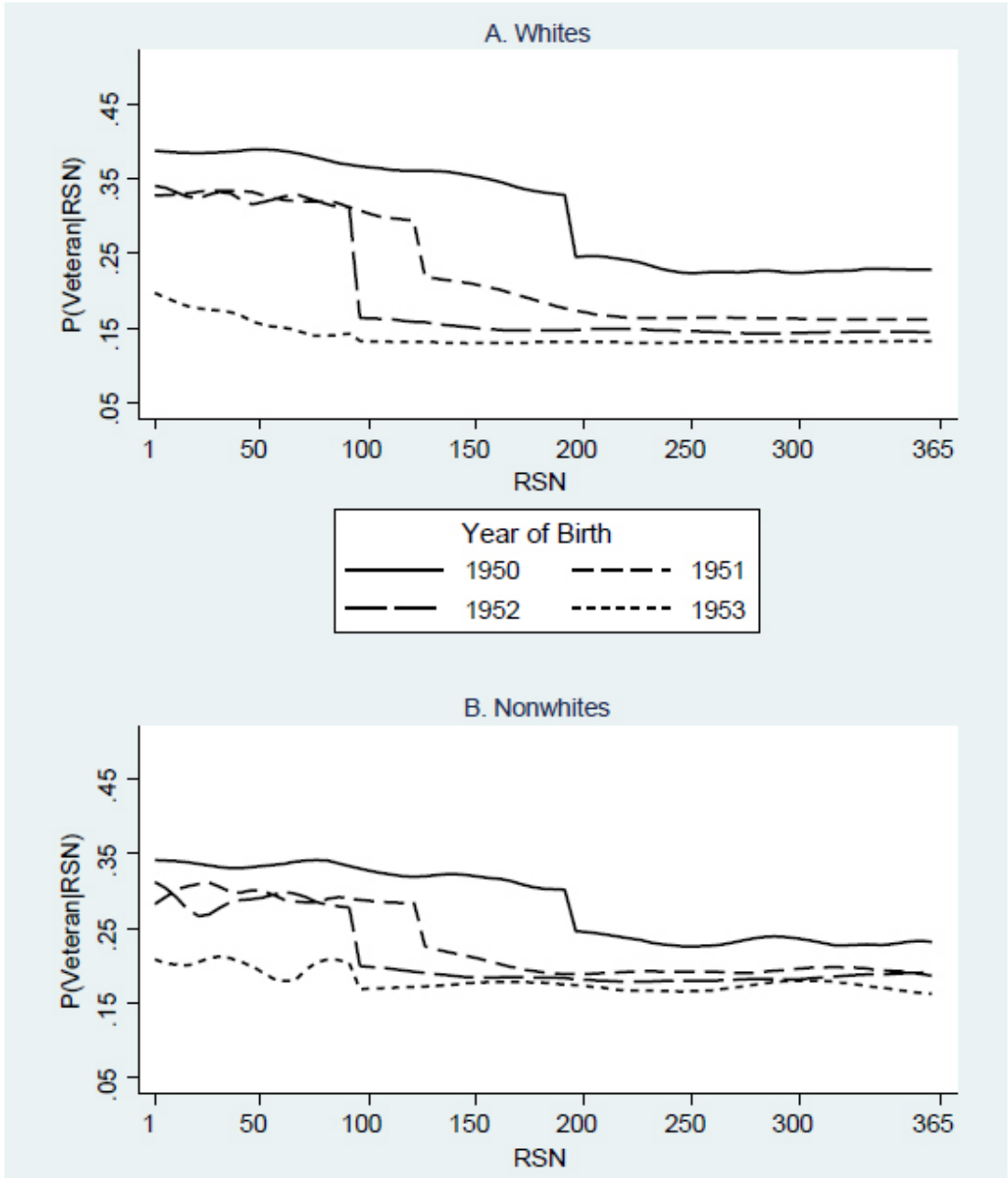


Figure 1. Draft-lottery Estimates of Vietnam-era Service Effects on ln(Earnings) for White Men Born 1950-52



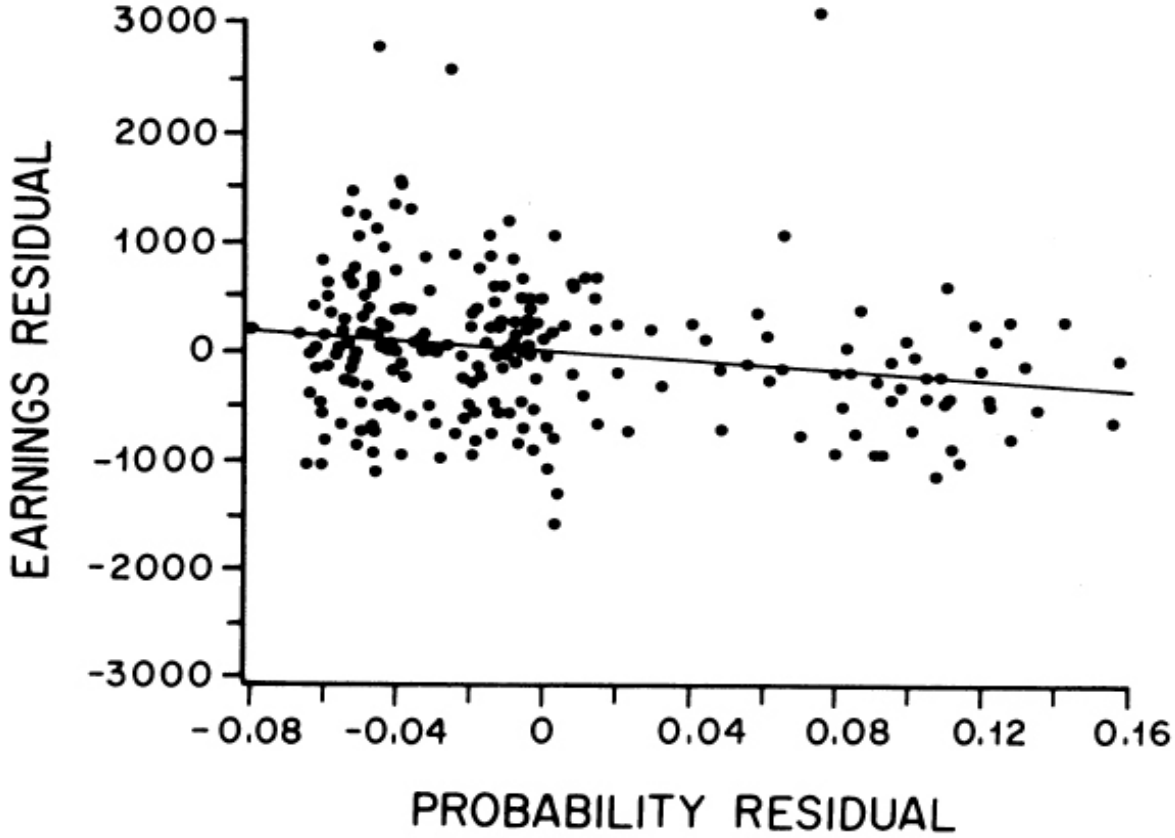


FIGURE 3. EARNINGS AND THE PROBABILITY OF VETERAN STATUS BY LOTTERY NUMBER

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

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

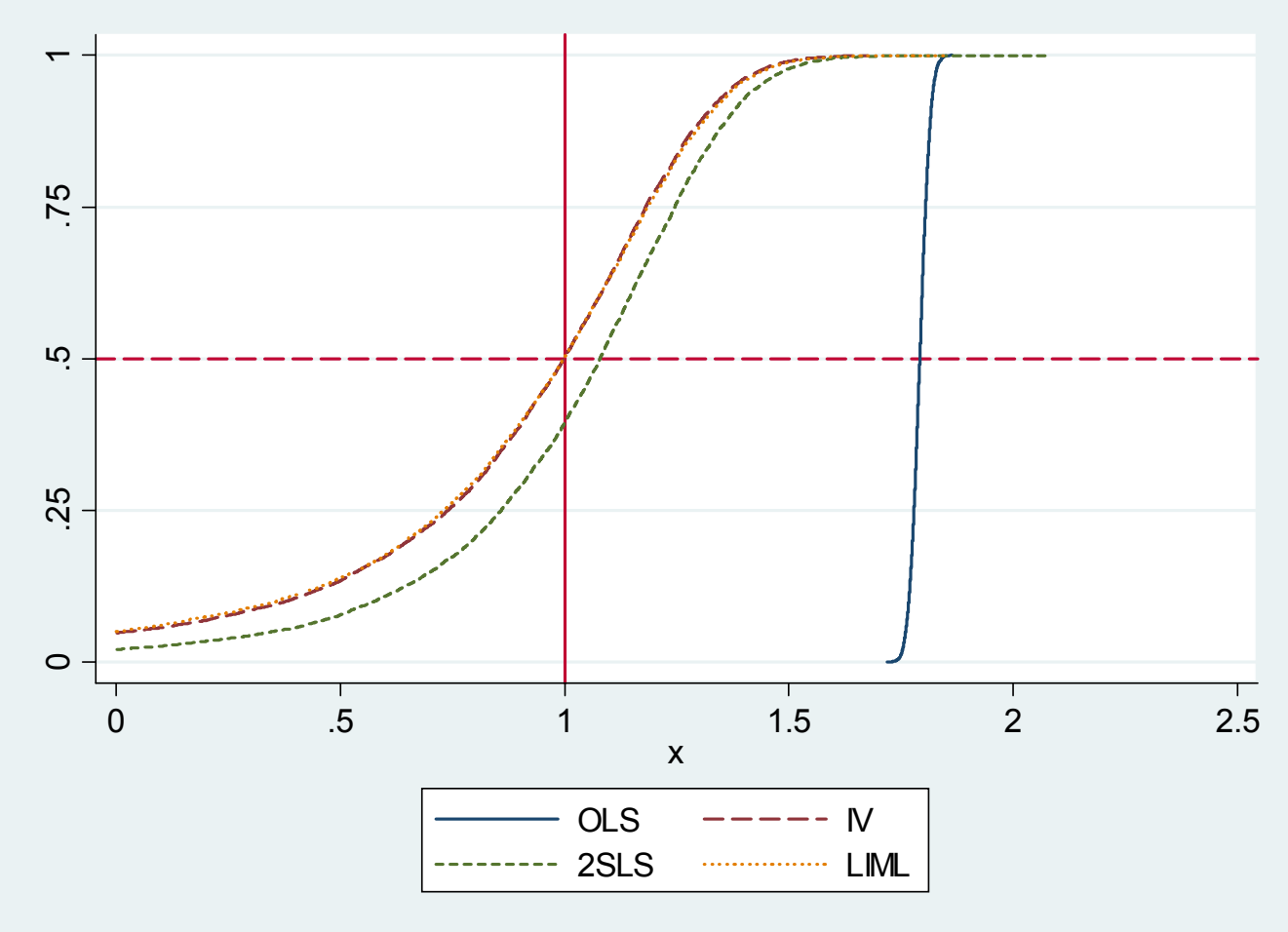


Figure 4.6.1: Distribution of the OLS, IV, 2SLS, and LIML estimators. IV uses one instrument, while 2S and LIML use two instruments.

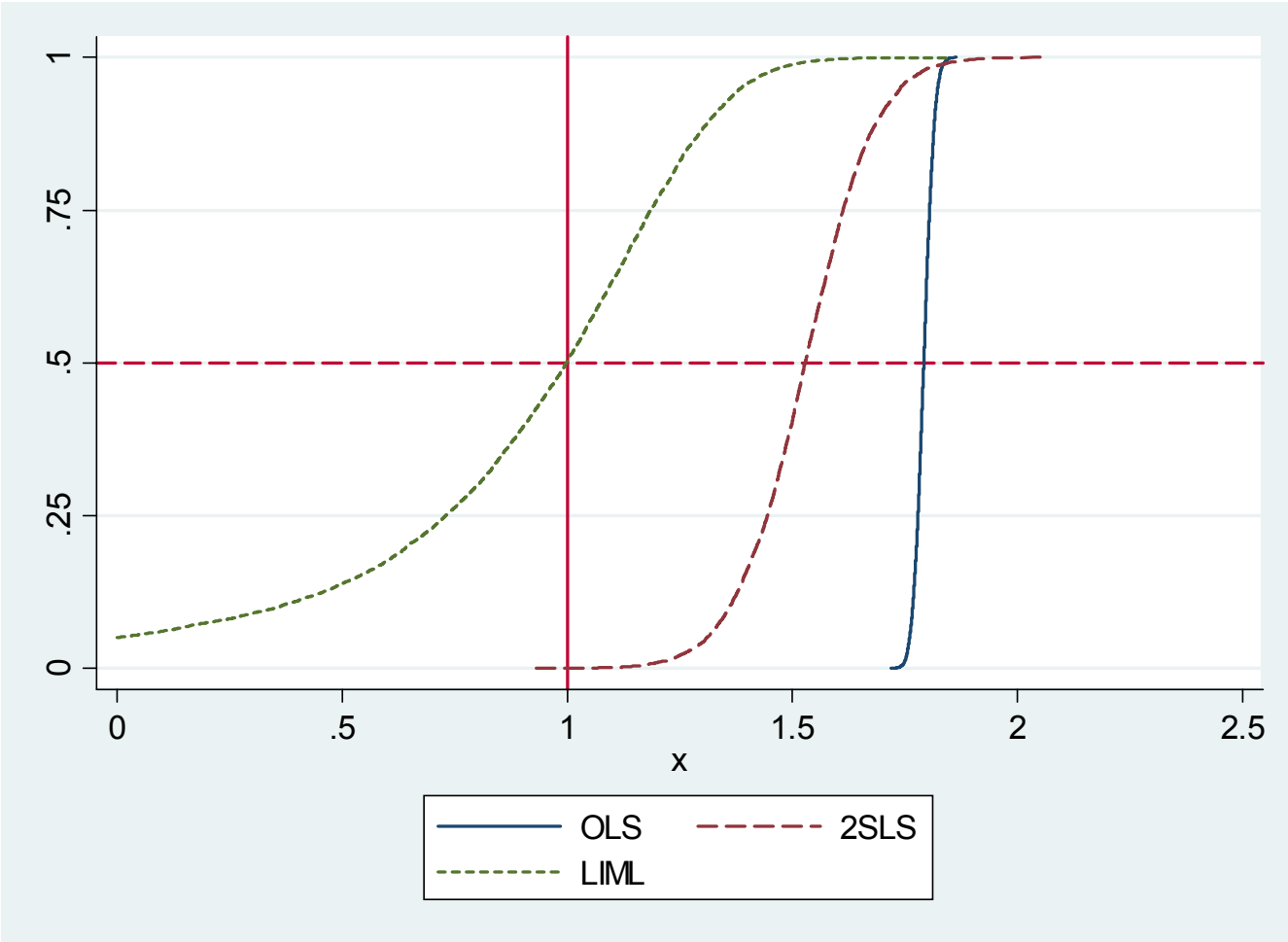


Figure 4.6.2: Distribution of the OLS, 2SLS, and LIML estimators with 20 instruments

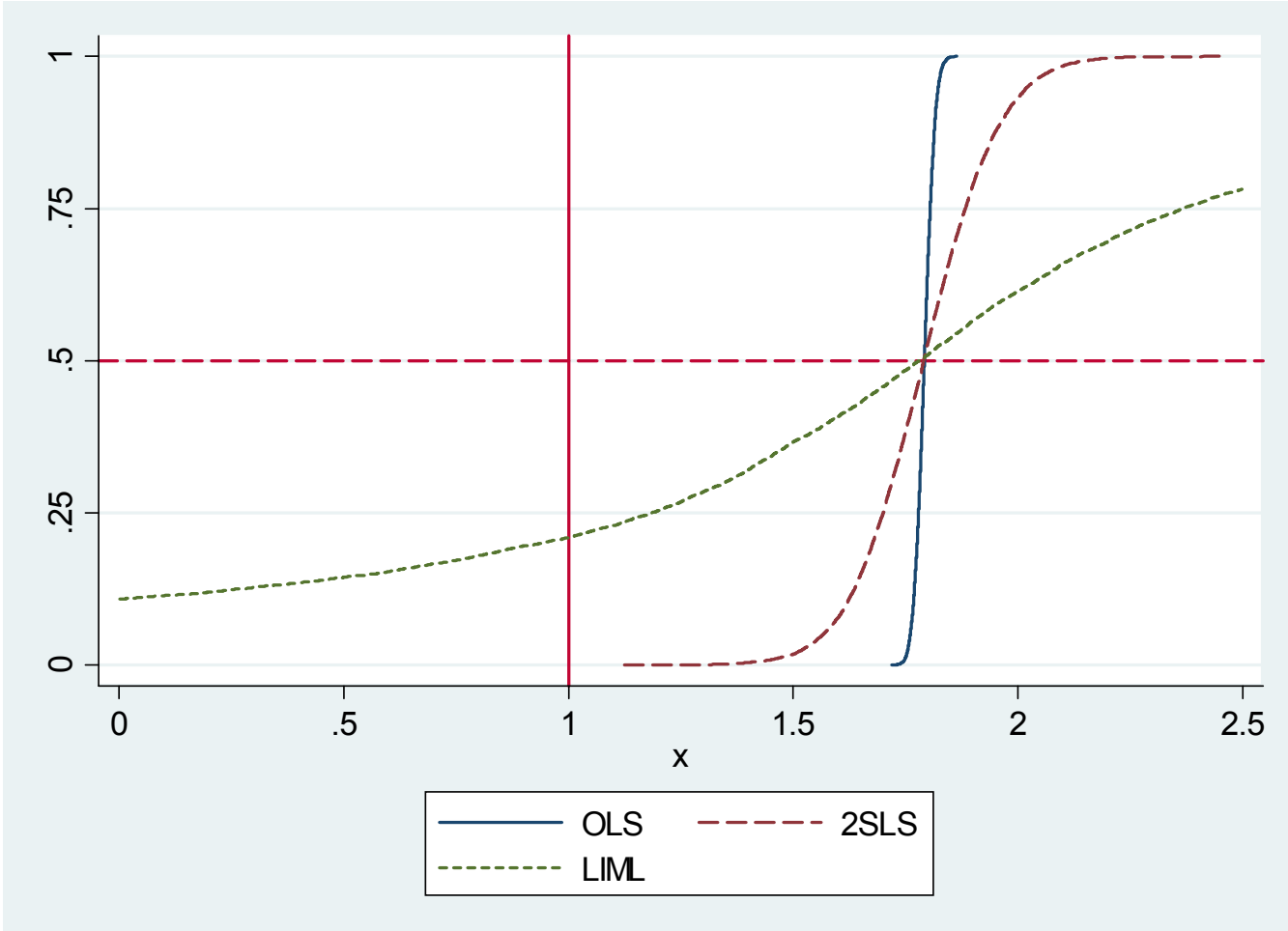


Figure 4.6.3: Distribution of the OLS, 2SLS, and LIML estimators with 20 worthless instruments

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.

14.387 Applied Econometrics: Mostly Harmless Big Data

Fall 2014