Econometric Methods - Instrumental Variables with Weak/Unidentified Instruments

Nicky Grant

November 2014

Overview of Lecture

1. Recap of 2SLS and assumptions
2. Properties of 2SLS with weak/unidentified instruments
3. Simulation evidence
4. Testing the identification assumption
5. Empirical application: Quarter of Birth as instrument for Educ. (Angrist & Krueger (1991))

2SLS Assumptions

2SLS1 $y_i = x_i'\beta_0 + u_i$

2SLS2 $E[u_i z_i] = 0$

2SLS3 $E[x_i z_i']$ is full rank and finite.

2SLS4 $E[u_i^2|Z] = \sigma^2$

2SLS5 $E[u_i u_j|Z] = 0$ for $i \neq j$

2SLS6 (x_i, z_i, u_i) is an i.i.d sequence

2SLS7 $E[x_i x_i']$, $E[z_i z_i']$ both finite and p.d

2SLS3 - instrument(s) correlated with endogenous variable(s).

2SLS known as the **Identification Condition**

2SLS3 inconsistent when 2SLS3 fails

If 2SLS3 almost fails- 2SLS estimator has high variance

Recap of Two Stage Least Squares (2SLS)

Two Stage Least Squares (2SLS) (define $Z = (z_1, .., z_n)'$)

1. Regress x_i on z_i save fitted value $\hat{x}_i := \hat{\Pi} z_i$ where $\hat{\Pi} = (Z'Z)^{-1}Z'X$
2. Regress y_i on $\hat{x}_i$

$P_Z = Z(Z'Z)^{-1}Z'$ then $\hat{X} = Z\hat{\Pi} = P_Z X$

$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (X'P_z X)^{-1}X'P_z y$

When $m = k$ $\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$

2SLS Estimator: Recap of Properties

$$\hat{\beta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

$$= \beta_0 + (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'u \text{ (by Linearity)}$$

$$= \beta_0 + (\frac{X'Z}{N}(\frac{Z'Z}{N})^{-1}\frac{Z'X}{N})^{-1}\frac{X'Z}{N}(\frac{Z'Z}{N})^{-1}\frac{Z'u}{N}$$

Define $Q_{ZZ} = E[z_i z_i']$, $Q_{XZ} = E[x_i z_i']$, $Q_{XP_ZX} = Q_{XZ}Q_{ZZ}^{-1}Q_{XZ}'$

Under 2SLS1-2SLS7

$\hat{\beta}_{2SLS} \xrightarrow{p} \beta_0$ **(Consistency)**

$\sqrt{N}(\hat{\beta}_{2SLS} - \beta_0) \xrightarrow{d} N(0, \sigma^2 Q_{XP_ZX}^{-1})$ **(Asymptotic Distribution)**

Violation of 2SLS3 in a Special Case

$$y = \beta_{00} + \beta_{01} x_{1i} + u_i$$

$x_i = (1, x_{i1})'$, $z_i = (1, z_{i1})'$

$$E[z_i x_i'] = \begin{pmatrix} 1 & E[x_{i1}] \\ E[z_{i1}] & E[z_{i1} x_{i1}] \end{pmatrix} \text{ full rank} \Rightarrow Cov(z_{i1}, x_{i1}) \neq 0$$

$Cov(z_{1i}, x_{1i}) = 0$ then z_{1i} doesn't explain variation in x_{1i}

If 2SLS3 fails then 2SLS cannot identify β_0 (unidentified).

2SLS is biased in this case

If $Cov(z_{i1}, x_{i1}) \neq 0$ but is small, higher the variance of 2SLS.

Often exog. instruments small/no correlation with endog. variables.

Properties of 2SLS in Unidentified case $m = k = 1$

$$y_i = \beta_0 x_i + u_i \quad \text{assume } \beta_{00} = 0 \text{ for simplicity}$$

$$\hat{\beta}_{2SLS} = (Z'X)^{-1}Z'y = \beta_0 + \frac{N^{-1}Z'u}{N^{-1}Z'X}$$

2SLS3 fails if $E[z_i x_i] = 0$ (since $z_i x_i$ is scalar)

If $E[x_i z_i] = 0$ then $N^{-1}Z'X = N^{-1}\sum_{i=1}^{N} z_i x_i \xrightarrow{p} 0$

By CLT $N^{-1/2}Z'X$ & $N^{-1/2}Z'u$ converge in dist. to some normal r.v

$\hat{\beta}_{2SLS}$ converges in dist. to β_0 plus some r.v

2SLS estimator is inconsistent

Properties of 2SLS in Unidentified case $m = k = 1$

$$N^{-1/2} Z'X = N^{-1/2} \sum_{i=1}^{N} z_i x_i \xrightarrow{d} N(0, E[(z_i x_i)^2]) \text{ 2SLS6 \& } E[x_i z_i] = 0$$

$$N^{-1/2} Z'u = N^{-1/2} \sum_{i=1}^{N} z_i u_i \xrightarrow{d} N(0, \sigma^2 E[z_i^2]) \text{ by 2SLS2,2SLS6.}$$

$$N^{-1/2} \sum_{i=1}^{N} \begin{pmatrix} z_i u_i \\ z_i x_i \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \zeta_{zu} \\ \zeta_{zx} \end{pmatrix}$$

Where ζ_{zu} and ζ_{zx} are two correlated Normal r.v

$$\hat{\beta}_{2SLS} \xrightarrow{d} \beta_0 + \frac{\zeta_{zu}}{\zeta_{zx}} \neq \beta_0 \quad \text{(2SLS inconsistent)} \tag{1}$$

The variance of 2SLS estimator in this case is infinite!

Comparison of Variance of 2SLS and OLS

Suppose 2SLS and OLS are consistent..Which estimator to prefer?

OLS1-6: $N^{1/2}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} N(0, \sigma^2 Q_{XX}^{-1})$

2SLS1-7 $N^{1/2}(\hat{\beta}_{2SLS} - \beta_0) \xrightarrow{d} N(0, \sigma^2 Q_{XP_ZX}^{-1})$

$\sigma^2 Q_{XX}^{-1}$ asy. var. of $N^{1/2}(\hat{\beta}_{OLS} - \beta_0)$ $AVar(N^{1/2}(\hat{\beta}_{OLS} - \beta_0))$

$\sigma^2 Q_{XP_ZX}^{-1}$ asy. var. of $N^{1/2}(\hat{\beta}_{OLS} - \beta_0)$ $AVar(N^{1/2}(\hat{\beta}_{2LS} - \beta_0))$

Can show $AVar(N^{1/2}(\hat{\beta}_{2LS} - \beta_0)) \geq AVar(N^{1/2}(\hat{\beta}_{OLS} - \beta_0))$

Comparison of Variance of 2SLS and OLS

Consider the simple case $x_i = (1, x_{1i})'$, $z_i = (1, z_{1i})'$

$$y_i = \beta_{00} + \beta_{01} x_{1i} + u_i$$

$$\boxed{AVar(N^{1/2}\hat{\beta}_{1,2SLS}) = \frac{\sigma^2 \sigma_{z1}^2}{Cov(z_{1i}, x_{1i})^2} = \frac{\sigma^2}{\rho^2 \sigma_{x1}^2}}$$

$\sigma_{z1}^2 = Var(z_{1i})$ $\sigma_{x1}^2 = Var(x_{1i})$, $\rho = Corr(z_{1i}, x_{1i})$

$$\boxed{AVar(N^{1/2}\hat{\beta}_{1,OLS}) = \frac{\sigma^2}{\sigma_{x1}^2}}$$

$\frac{AVar(N^{1/2}\hat{\beta}_{1,2SLS})}{AVar(N^{1/2}\hat{\beta}_{1,OLS})} = 1/\rho^2$ (see exc. 7)

Simulation: Properties of 2SLS with weak/unidentification

$$y_i = 1 + x_{1i} + u_i \tag{2}$$

$$x_i = 1 + \pi_{01} z_{1i} + v_i \tag{3}$$

$Corr(v_i, u_i) = 0.4$, $Cov(z_{1i}, u_i) = 0$

$(x_{1i}, z_{1i}, u_i, v_i)$ are all i.i.d $N(0,1)$ for $i = 1, .., N$.

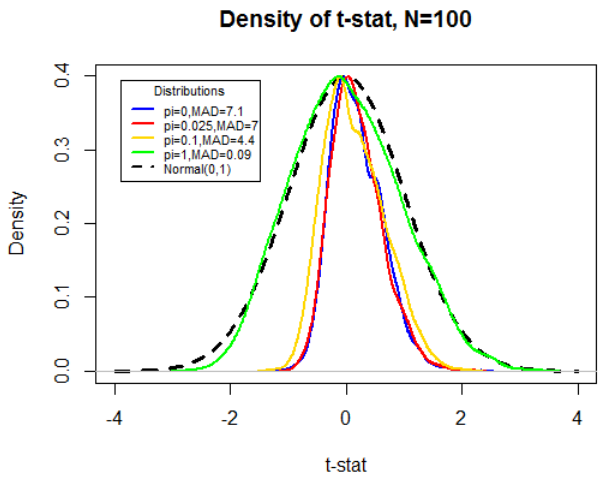
$\pi_{01} = \{0, 0.025, 0.1, 1\}$ for $N = \{100, 500, 1000, 5000, 10000, 50000\}$.

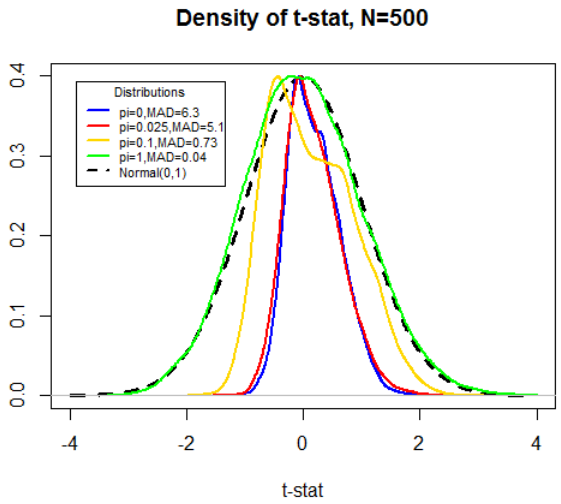
Plot density of $t - stat = \frac{\hat{\beta}_{2SLS} - 1}{s.e(\hat{\beta}_{2SLS})}$ for $H_0 : \beta_{01} = 1$

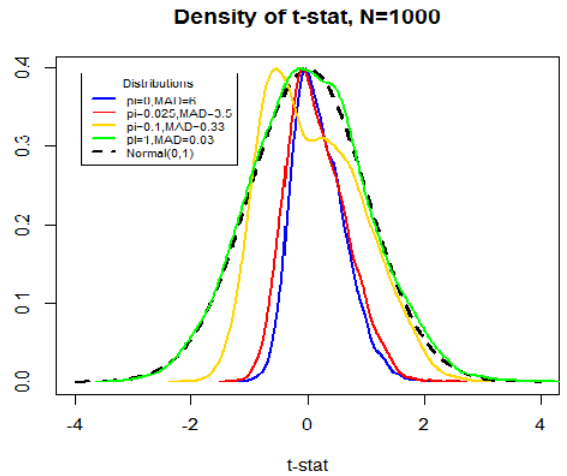
Smaller is π_{01} the poorer the normal approx to t-stat

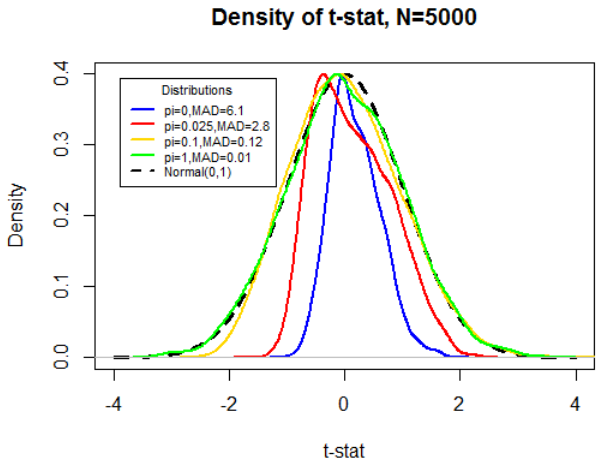
Asy. variance increases as π_{01} decreases.

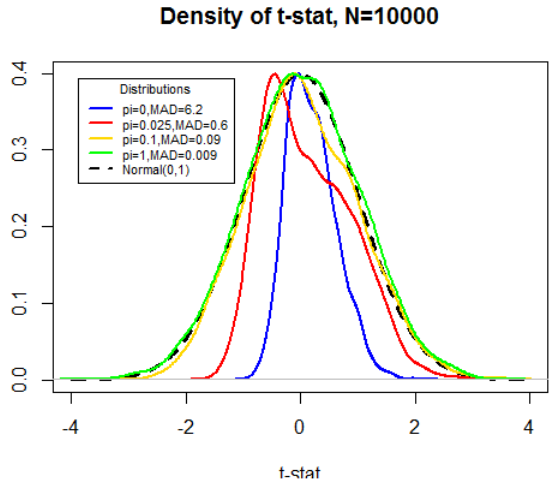
Simulation Results

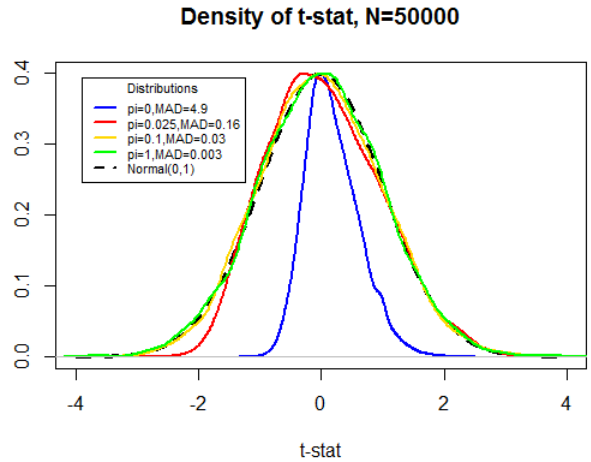












MAD is the Monte Carlo estimate of the Mean Absolute Deviation of $\hat{\beta}_{1,2SLS}$ ($E[|\hat{\beta}_{1,2SLS} - \beta_{01}|]$)

Testing Identification: 1 Endogenous regressor, 1 Instrument, No controls

We can test whether $E[x_i z_i']$ is full rank

$$y_i = \beta_{00} + \beta_{01} x_{1i} + u_i \tag{4}$$

$$x_i = \pi_{00} + \pi_{01} z_{1i} + v_i \tag{5}$$

$det(E[x_i z_i']) = Cov(x_{1i}, z_{1i})$

$Cov(x_{1i}, z_{1i}) = \pi_{01} \sigma^2_{z1}$

$E[x_i z_i']$ full rank equivalent to $\pi_{01} \neq 0$

Estimate π_{01} by OLS in (5) and perform a t-test $\pi_{01} = 0$

Testing Identification: 1 Endogenous regressor, 1 Instrument, with controls

$$y_i = \beta_{00} + \beta_{01} x_{1i} + \beta_w' w_i + u_i \tag{6}$$

$$x_i = \pi_{00} + \pi_{01} x_{1i} + \pi_w' w_i + v_i \tag{7}$$

w_i is a $(k-2) \times 1$ vector of exogenous control variables,
z_{1i} is an instrument for the endogenous x_{1i}.

$x_i = (1, x_{1i}, w_i')'$, $z_i = (1, z_{1i}, w_i')'$

We can show $E[x_i z_i']$ is full rank if $\pi_{01} \neq 0$ **(See exc. 7)**

Test 2SLS by t-test that $\pi_{01} = 0$ using OLS in (7).

Testing Identification: 1 Endogenous regressor, Many instrument, with controls

$$y_i = \beta_{00} + \beta_{01} x_{1i} + \beta_w' w_i + u_i \tag{8}$$

$$x_i = \pi_{00} + \pi_{01}' z_{1i} + \pi_w' w_i + v_i \tag{9}$$

w_i is a $(k-2) \times 1$ vector of exogenous control variables,
z_{1i} is an $m_1 \times 1$ instrument for the endogenous scalar x_{1i}.

$x_i = (1, x_{1i}, w_i')'$, $z_i = (1, z_{1i}', w_i')'$

We can show $E[x_i z_i']$ is full rank if at least one elements of π_{01} does not equal zero. **(See exc. 7)**

Test 2SLS3 by joint F-Test that $\pi_{01} = 0$.

Testing Identification: General Case

Can allow many endogenous regressors also and test 2SLS3 by OLS

Will not cover for simplicity- most papers have 1 endog. variable

F-Stat of the test $\pi_{01} = 0$ used as measure of instrument 'strength'

Higher the F-stat the less likely 2SLS unidentified

Rule of thumb: $F > 10$ for 2SLS not to have poor properties (Staiger & Stock (1997))

Angrist & Krueger (1991):Returns to Schooling.

$$w_i = \alpha_0 + \beta_0 s_i + \beta_w' w_i + u_i \tag{10}$$

w_i =wages,s_i =yrs of schooling, w_i control variables including dummy for race, marriage status, urban, year of birth, state of birth.

Use QoB (and interaction with year dummies) as instruments for w_i

Child must turn 6 by January 1 in the year they start school

Minimum school leaving age is 16

Children born later in year start school slightly older

If children born earlier in year don't compensate by staying longer leads to differences in years of schooling

Quarter of Birth and Total Years of Education

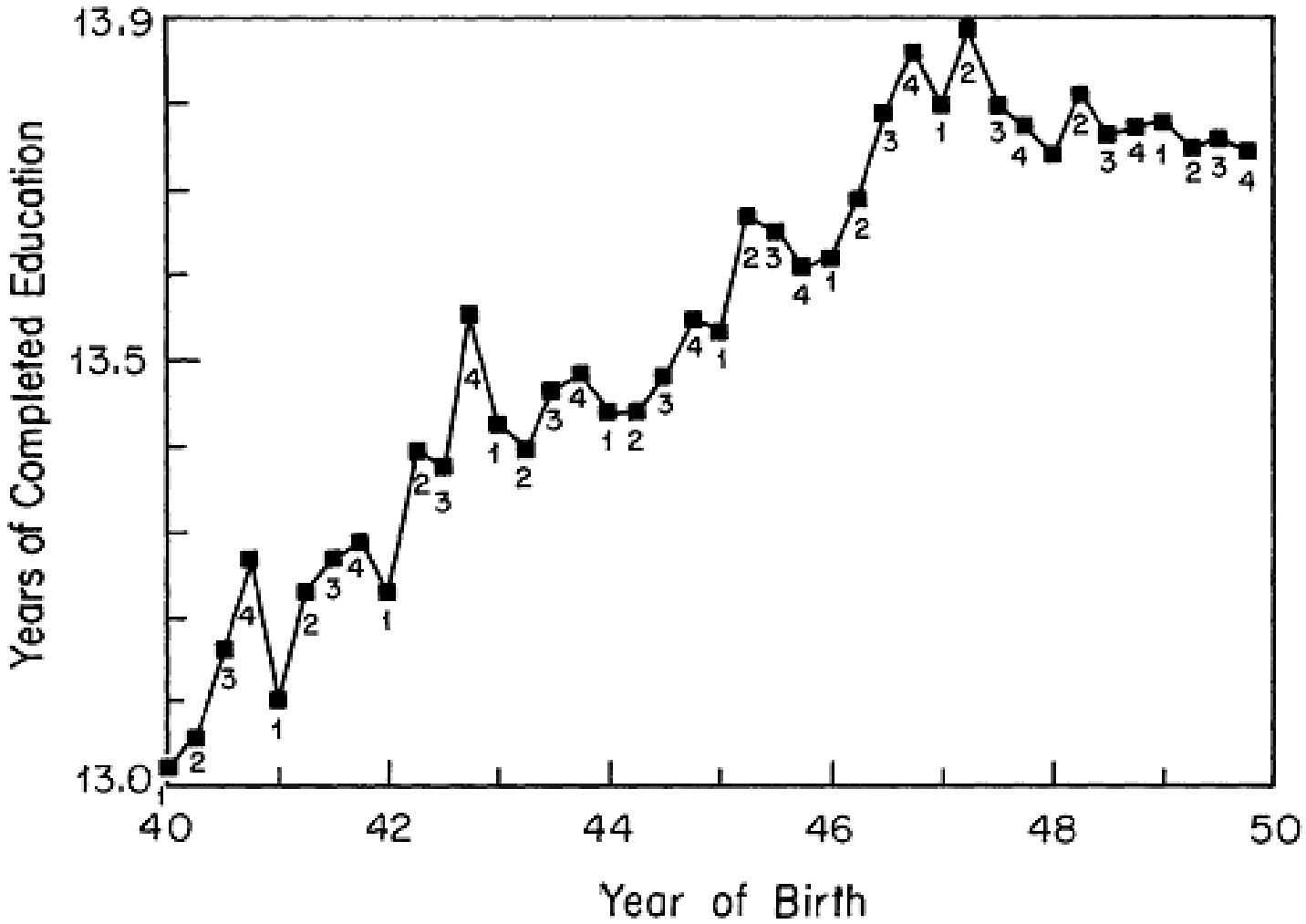


FIGURE II
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

AK1991: OLS/2SLS Estimators (1930-1939 cohort)

OLS and TSLS Estimates of the Return to Education for Men Born 1930–1939: 1980 Census[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0711	0.0891	0.0711	0.0760	0.0632	0.0806	0.0632	0.0600
	(0.0003)	(0.0161)	(0.0003)	(0.0290)	(0.0003)	(0.0164)	(0.0003)	(0.0299)
Race (1 = black)	—	—	—	—	−0.2575	−0.2302	−0.2575	−0.2626
					(0.0040)	(0.0261)	(0.0040)	(0.0458)
SMSA (1 = center city)	—	—	—	—	0.1763	0.1581	0.1763	0.1797
					(0.0029)	(0.0174)	(0.0029)	(0.0305)
Married (1 = married)	—	—	—	—	0.2479	0.2440	0.2479	0.2486
					(0.0032)	(0.0049)	(0.0032)	(0.0073)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	—	—	−0.0772	−0.0801	—	—	−0.0760	−0.0741
			(0.0621)	(0.0645)			(0.0604)	(0.0626)
Age-squared	—	—	0.0008	0.0008	—	—	0.0008	0.0007
			(0.0007)	(0.0007)			(0.0007)	(0.0007)

AK1991: OLS/2SLS Estimators (1930-1939 cohort) with State Controls

OLS and TSLS Estimates of the Return to Education for Men Born 1930–1939: 1980 Census[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0673	0.0928	0.0673	0.0907	0.0628	0.0831	0.0628	0.0811
	(0.0003)	(0.0093)	(0.0003)	(0.0107)	(0.0003)	(0.0095)	(0.0003)	(0.0109)
Race (1 = black)	—	—	—	—	−0.2547	−0.2333	−0.2547	−0.2354
					(0.0043)	(0.0109)	(0.0043)	(0.0122)
SMSA (1 = center city)	—	—	—	—	0.1705	0.1511	0.1705	0.1531
					(0.0029)	(0.0095)	(0.0029)	(0.0107)
Married (1 = married)	—	—	—	—	0.2487	0.2435	0.2487	0.2441
					(0.0032)	(0.0040)	(0.0032)	(0.0042)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
50 State-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age	—	—	−0.0757	−0.0880	—	—	−0.0778	−0.0876
			(0.0617)	(0.0624)			(0.0603)	(0.0609)
Age-squared	—	—	0.0008	0.0009	—	—	0.0008	0.0009
			(0.0007)	(0.0007)			(0.0007)	(0.0007)

AK (1991): Weak Instrument Problem

QoB likely exogenous but leads to very small changes in educ.

Orignal AK (1991) did not test the identification assumption

Bound, Baker & Jaeger (1995) show F stat is small, **accepting the null that instrument not identified**

Simulate random QoB data and use as instrument and find similar results to AK(1991)!

AK(1991) results likely dubious

Recap of Lecture

2SLS inconsistent when identification assumption fails (unidentified).

If identified but small correlation between instruments and endog variance- 2SLS has a high variance

Instruments commonly weak or unidentified

Very hard to find exogenous variable with a strong correlation with endog. variables

Reading

Greene chapter 8.7, 10.6.6

Does Compulsory School Attendance and Earnings?, Angrist & Krueger (1991), Quarterly Journal of Economics.
http://web.stanford.edu/~pista/angrist.pdf

Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak., Bound, Jaeger & Baker (1995), Journal of the American Statistical Association.
https://faculty.fuqua.duke.edu/~charlesw/s591/Methods/c04_Ashish/boundetal_probswinstrum_jasa95.pdf,
Extra Reading
Avoiding Invalid Instruments and Coping with Weak Instruments, Murray (2006), Journal of Economic Perspectives.
http://faculty.smu.edu/Millimet/classes/eco6374/papers/murray.pdf