# IV Estimation and its Limitations:

# Weak Instruments and Weakly Endogeneous Regressors

Laura Mayoral

IAE, Barcelona GSE and University of Gothenburg

Gothenburg, May 2015

# Roadmap

Deviations from the standard framework:

- Irrelevant and weak instruments

- Endogeneous and weakly endogeneous regressors

## 3. Violations to the relevance condition: set up

- Consider the following model

$$
\begin{aligned}
y &= X\beta + \epsilon; && (1)\\
X &= Z\Pi + v; && (2)\\
corr(\epsilon, v) &\neq 0 && (3)
\end{aligned}
$$

Remarks:

- X is endogenous since $corr(\epsilon, v) \neq 0$
- Unless otherwise stated it's assumed throughout that Z is exogenous $(E(Z'\epsilon) = 0)$
- Without loss of generality, we omit exogenous regressors (if they exist, they can be partialled out)

## 3.1. Irrelevant instruments

- Recall that if there is just one IV then

$$\hat{\beta}_{2sls} = \beta + \frac{\widehat{cov(Z_i, y)}}{\widehat{cov(Z_i, X_i)}}$$

- If $cov(Z_i, X_i) = 0 \longrightarrow$ Z is irrelevant.

- General case (more instruments): if $E(Z'X) = 0 \longrightarrow$ Z is irrelevant.

- Using the notation above: $\Pi = 0 \longrightarrow$ Z is irrelevant

## Irrelevant instruments, II

- What happens when Z is irrelevant?

- $\hat{\beta}_{2sls}$ is not identified.

- $\hat{\beta}_{2sls}$ is inconsistent (we knew this already!)

- The distribution of $(\hat{\beta}_{2sls} - \beta)$ is Cauchy-like

- The bias of $\hat{\beta}_{2sls}$ tends to that of $\hat{\beta}_{ols}$: The distribution of $\hat{\beta}_{2sls}$ is centered around the $plim(\hat{\beta}_{ols})$.

## 3.2. Weak instruments

- Recall that if there is just one IV then

$$\hat{\beta}_{2sls} = \beta + \frac{\widehat{cov(Z_i, y)}}{\widehat{cov(Z_i, X_i)}}$$

- If $cov(Z_i, X_i) > 0$ but close to zero: Z is weak.

- Why? $\widehat{cov(Z_i, X_i)}$ would be close to zero $\longrightarrow$ the bias of $\hat{\beta}_{2sls}$ will be very large!

- In fact, the main problem derives from the fact that the finite-sample distribution is very different from the asymptotic one (and remember that 2SLS's justification is asymptotic!).

# Weak instruments, consequences

- Although strictly speaking the conditions for consistency of $\hat{\beta}_{2sls}$ are met (since $E(X'Z) \neq 0$) standard asymptotics yield very poor approximation to the finite-sample distributions when instruments are weak.

- As a result $\hat{\beta}_{2sls}$ is weakly identified, (i.e., its distribution is not well approximated by their standard asymptotic distribution).

- The source of the problem is not small-sample problems in a conventional sense, but rather, limited information, (see Bound et al. 1995).

- To evaluate how severe is the weak instrument problem, we need to have analytical expressions that approximate better the finite sample properties of the estimators when instruments are weak.

## Analytical approximations to the finite sample distributions of IV estimators when instruments are weak

- Different approaches:

- Approach 1: Assume errors are normal and Z is fix. Then we can derive the exact distribution

- Define:

$$\mu^2 = \Pi' Z' Z \Pi / \sigma_v^2$$

- $\mu^2$: concentration parameter, is a measure of the strength of the instruments.

(In particular: it measures the share of the variance of X explained by Z –normalised by the variance of $v$–).

- Nelson and Startz (1990) show that

$$\mu(\hat{\beta}_{2sls} - \beta) = \Lambda$$

- $\mu$ plays the role that is usually played by $\sqrt{N}$ (sample size).

- If $\mu \xrightarrow{p} \infty$: instruments are strong, $\Lambda$ is the usual normal distribution

- If $\mu$ is small: then $\Lambda$ is not standard and estimates will be very biased.

- How big is the bias? VERY big if $\mu$ is small.
- The following graph plots $\Lambda$ for different values of $\mu$
- In this example: $\beta = 0$, $\rho = .95$ is the bias of the OLS estimator.

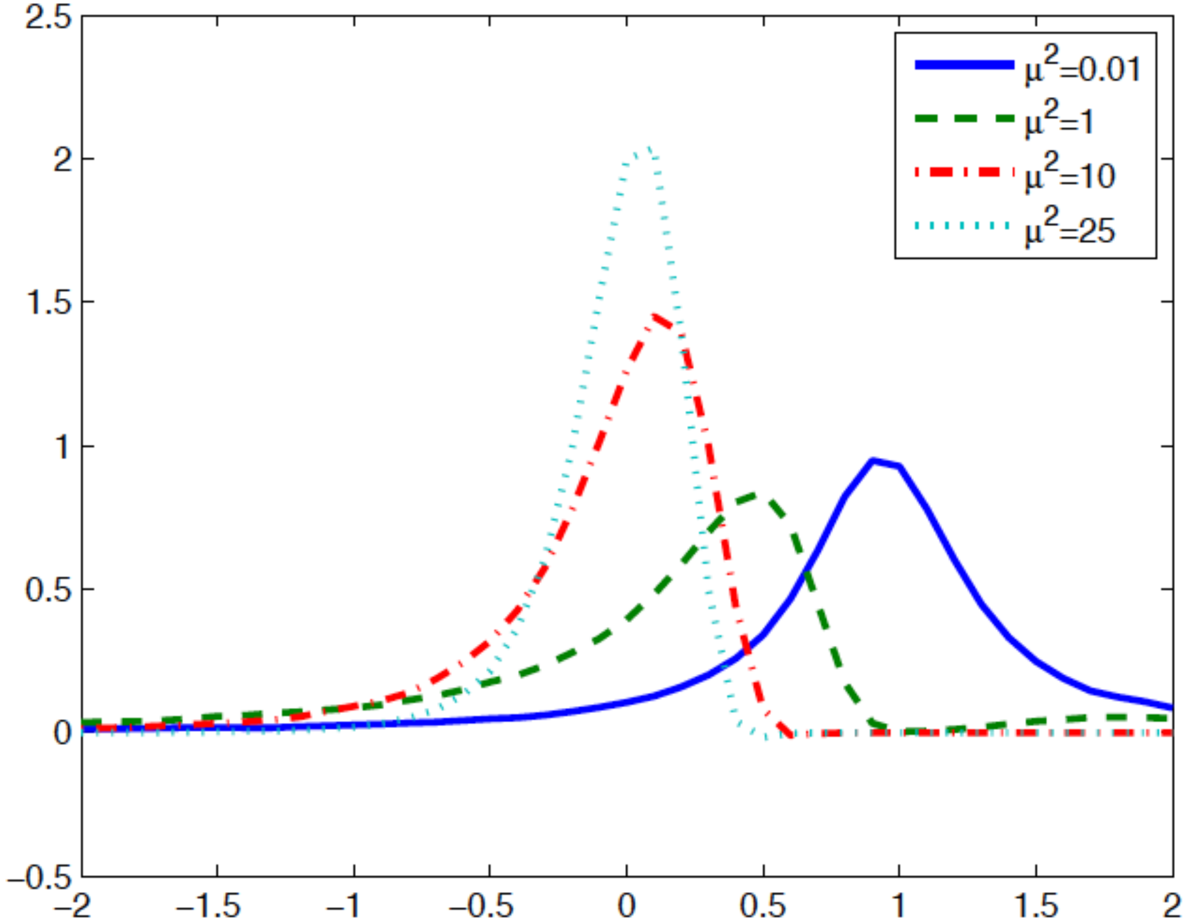



Figure 1: Finite-sample distribution of the TSLS estimator given by formula (4) fo
different values of the concentration parameter. $\rho = 0.95$, $\frac{\sigma_\epsilon}{\sigma_v} = 1$, $\beta_0 = 0$, $r = 1$

# Approach 2: Weak instrument asymptotics

- Approach 1 is developed under very demanding assumptions: normality, fixed Z.

- Staiger and Stock (1997): showed that very similar results can be obtained (i.e., same distribution of $\beta$, $\Lambda$) under general conditions (Z not fixed, errors non normal).

# Approach 2: Weak instrument asymptotics, II

- Their approach: assume $\Pi = \Pi_N = C/\sqrt(N)$ such that

$$\begin{aligned} y &= X\beta_0 + \epsilon; && (4) \\ X &= Z\Pi + v; && (5) \\ \Pi &= \Pi_N = C/\sqrt(N) && (6) \\ corr(\epsilon, v) &\neq 0 && (7) \end{aligned}$$

estimate $\beta$ and compute the asymptotic distribution of $\beta$ under $\Pi = \Pi_N = C/\sqrt(N)$

$\longrightarrow$ weak instrument asymptotics.

## Weak instrument asymptotics

- As mentioned before, weak instruments should not be thought of a finite sample problem!

- Staiger and Stock show that:

  - for each sample size (even for a very large one) there will exist some values of correlation between the instrument and the regressor such that the quality of normal approximation is poor

## What's the meaning $\Pi = \Pi_N = C/\sqrt(N)$?

- This is a just 'trick' to obtain analytical expressions that approximate the distribution in finite samples in a better way. That is, we don't truly believe that $\Pi = \Pi_N = C/\sqrt(N)$, but it is useful to assume this setup.

- Why?

- We saw before that for any $\Pi$ arbitrarily close to zero (but $> 0$), standard asymptotic theory will not be informative as the limited information problem will be eventually overcome by an infinite sample size. This solution is not satisfactory as we never have such a thing!

- By choosing $\Pi = \Pi_N = C/\sqrt(N)$, $\mu^2$ remains constant as the sample size increases (rather than going to $\infty$ as it would for a fix $\Pi > 0$) (so the weak instruments problem remains even if $N \longrightarrow \infty$)

## Summary of weak IV asymptotic results (Staiger and Stock, 1997).

- $\hat{\beta}_{2sls}$ is not consistent and non-normal

- The analytical expressions obtained under weak IV asymptotics: provide very good approximations of the finite sample distributions when the correlation between the instruments and the endogenous regressor is small.

- Test statistics (including J-test of overindentifying restrictions) do not have standard distributions.

# Summary of weak IV asymptotic results, II.

- Bias of $\hat{\beta}_{1_{2sls}}$ :

- If Z is irrelevant (E(Z'X)=0) $\longrightarrow$ $\hat{\beta}_{2sls}$ is centered around the plim($\hat{\beta}_{ols}$)

Note. Remember $[plim\hat{\beta}_{ols} = \beta + E(X'X)^{-1}E(X'\epsilon)]$

- If Z is weak: the bias of $\hat{\beta}_{2sls}$ tends to the bias of $\hat{\beta}_{ols}$

# 4. Violations to the exclusion condition: Set up

- Let's now consider violations to the exclusion restriction

- Set up

$$y = X\beta + \epsilon; \quad (8)$$
$$X = Z\Pi + v; \quad (9)$$
$$\epsilon = Z\gamma + \epsilon^* \quad (10)$$
$$corr(\epsilon^*, v) \neq 0 \quad (11)$$

- Unless otherwise stated, assume Z is strong i.e. $\Pi \neq 0$ and is large.

- If $\gamma \neq 0 \longrightarrow$ Z is endogenous, (exclusion restriction is violated)

# Endogeneous instruments, consequences

- $\beta$ is not consistent.

- The bias (1 endogenous variable, 1 instrument):

$$\hat{\beta}_{2sls} - \beta \xrightarrow{p} \frac{cov(Z, \epsilon)}{cov(Z, X)} = \frac{\gamma}{\Pi}$$

- Very important:

- Any correlation between Z an $\epsilon$ will be magnified if the correlation between Z and X is small (i.e., if $\Pi$ is small)

## An example: Estimating the causal effect of years of education on lifetime earnings, Angrist and Krueger (1991).

- Education is likely to be endogenous. Why?

- Omitted variable: innate ability.

  - More talented people will tend to remain in school longer.

  - Also, more talented people will tend to earn more money.

- Angrist and Krueger (1991)'s approach: use "quarter of birth" as IV.

- Their argument: US compulsory schooling laws are in terms of age, not number of years of schooling completed. If compulsory schooling age is 16, you can drop out on your 16th birthday (even if in the middle of the school year).

- School entry is once a year, and cutoffs are based on birthdays. Beginning school: kids that are 6 years olds by Sep. 1st

- The combination of these two generates variation in schooling for those who drop out as soon as they can.

- Consider two students, one born on August 31st and another born on Sep. 1st: by the time they turn 16, the second has ha 1 additional year of education.

# Instrument validity

- Do birthdays satisfy the exclusion restriction, or could birthdays be correlated with earnings for other reasons than their effect on schooling?
  - Birthday affects, e.g., age rank in class?.
- Do birthdays indeed affect schooling? $\longrightarrow$ Check the first stage.

# Angrist and Krueger data

- Huge dataset.

Data are from the 1980 US Census. 329,509 men born 1930 to 1939 (i.e. in their 40s when observed). For these men we have year of birth, quarter of birth, years of schooling, and earnings in 1979.

## Do birthdays satisfy the exclusion restriction?,

- Bound, Jaeger and Baker (1995): argue that maybe Z is not exogenous:

  - Some evidence that quarter of birth is related to school attendance rates, performance in reading, maths, etc.

  - Differences in physical and mental health of individuals born in different times of the year.

- Key point:

  - Although the correlation between Z and $\epsilon$ is likely to be small, it gets magnified by the very small correlation between quarter of birth and education!!.

  - As a result, the bias can be very large.

- Bound et al: used Angrist and Krueger data but instead of using the actual quarter of birth, they randomly assigned a quarter of birth to each observation.

- This 'random' quarter of birth is exogenous but also, totally uncorrelated with education.

- However, the results using this instrument were very similar than those obtained with the true quarter of birth!

- Nothing in the second stage regression could suggest that the new Z was totally irrelevant!

- Another important point: the weak instrument problem was important even if the sample size was huge!!

# Many weak instruments

- Angrist and Krueger used as instruments the interactions between quarter of birth and year of birth and quarter of birth and state of birth.

- Adding many weak instruments makes the problem even worse: the F statistic of the first stage gets smaller $\longrightarrow$the bias gets worse.

- The next three tables are from Bound et al..

  - The first table shows that by increasing the number of (weak) instruments, the F statistic that tests the joint significance of the instruments in the the first stage decreases.

*Table 1. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings (standard errors of coefficients in parentheses)*

| | (1) OLS | (2) IV | (3) OLS | (4) IV | (5) OLS | (6) IV |
|---|---|---|---|---|---|---|
| Coefficient | .063<br>(.000) | .142<br>(.033) | .063<br>(.000) | .081<br>(.016) | .063<br>(.000) | .060<br>(.029) |
| $F$ (excluded instruments) | | 13.486 | | 4.747 | | 1.613 |
| Partial $R^2$ (excluded instruments, ×100) | | .012 | | .043 | | .014 |
| $F$ (overidentification) | | .932 | | .775 | | .725 |
| *Age Control Variables* | | | | | | |
| Age, Age$^2$ | x | x | | | x | x |
| 9 Year of birth dummies | | | x | x | x | x |
| *Excluded Instruments* | | | | | | |
| Quarter of birth | | x | | x | | x |
| Quarter of birth × year of birth | | | | x | | x |
| Number of excluded instruments | | 3 | | 30 | | 28 |

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), and 8 Regional dummies as control variables. $F$ (first stage) and partial $R^2$ are for the instruments in the first stage of IV estimation. $F$ (overidentification) is that suggested by Basmann (1960).

- Table 2: many weak instruments –contains instruments based on quarter of birth interacted with state of birth and year of birth.

- Table 3: uses the randomly generated quarters of birth (i.e., all instruments are irrelevant).

- Compare Tables 1/2 with 3: just by looking at the second stage regression you won't be able to detect that instruments are irrelevant: both tables look similar!

- but the first stage shows that the regressions are problematic– look at the F's!

- Notice that the OLS estimates and those obtained with irrelevant instruments are very similar, as the theory predicts.

*Table 2. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings, Controlling for State of Birth (standard errors of coefficients in parentheses)*

| | *(1) OLS* | *(2) IV* | *(3) OLS* | *(4) IV* |
|---|---|---|---|---|
| Coefficient | .063 | .083 | .063 | .081 |
| | (.000) | (.009) | (.000) | (.011) |
| *F* (excluded instruments) | | 2.428 | | 1.869 |
| Partial $R^2$ (excluded instruments, ×100) | | .133 | | .101 |
| *F* (overidentification) | | .919 | | .917 |
| *Age Control Variables* | | | | |
| Age, Age$^2$ | | | x | x |
| 9 Year of birth dummies | x | x | x | x |
| *Excluded Instruments* | | | | |
| Quarter of birth | | x | | x |
| Quarter of birth × year of birth | | x | | x |
| Quarter of birth × state of birth | | x | | x |
| Number of excluded instruments | | 180 | | 178 |

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), 8 Regional dummies, and 50 State of Birth dummies as control variables. *F* (first stage) and partial $R^2$ are for the instruments in the first stage of IV estimation. *F* (overidentification) is that suggested by Basmann (1960).

*Table 3. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings, Using Simulated Quarter of Birth (500 replications)*

| *Table (column)* | *1 (4)* | *1 (6)* | *2 (2)* | *2 (4)* |
|---|---|---|---|---|
| | | *Estimated Coefficient* | | |
| Mean | .062 | .061 | .060 | .060 |
| Standard deviation of mean | .038 | .039 | .015 | .015 |
| 5th percentile | −.001 | −.002 | .034 | .035 |
| Median | .061 | .061 | .060 | .060 |
| 95th percentile | .119 | .127 | .083 | .082 |
| | | *Estimated Standard Error* | | |
| Mean | .037 | .039 | .015 | .015 |

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509.

# Weakly endogeneous regressors

- In practice, it is very likely that Z and $\epsilon$ have some correlation.

- Thus, as before, it is useful to study local violations to the exclusion restriction

- Set up

$$
\begin{aligned}
y &= X\beta + \epsilon; && (12)\\
X &= Z\Pi + v; && (13)\\
\epsilon &= Z\gamma + \epsilon^* && (14)\\
\gamma &= \gamma_n = B/\sqrt{(N)} && (15)\\
corr(\epsilon^*, v) &\neq 0 && (16)
\end{aligned}
$$

- Interpretation:

- We don't believe that $\gamma$ is as described,

- this is just a trick to find analytical expressions to the finite sample distributions under mild violations of the endogeneity restriction (and it works very well!).

# Summary of asymptotic results under endogeneity of Z

1. Recall that $\gamma \neq 0 \longrightarrow \hat{\beta}_{2sls}$ is not consistent

2. But if $\gamma = \gamma_n = B/\sqrt{(}N)$, $\hat{\beta}_{2sls}$ is consistent!

3. Asymptotic distribution of $\hat{\beta}_{2sls}$ is normal with the same variance-covariance matrix but centered on a 'wrong' value.

$$T^{1/2}(\hat{\beta}_{2sls} - \beta_0) \xrightarrow{d} N(\frac{B}{\Pi}, \frac{\sigma_\epsilon^2}{\Pi^2 E(Z'Z)}) \tag{17}$$

## Some implications

- Under mild violations of the exclusion restriction: point estimates might still be ok! (IF sample size is large, estimates are strong and violation is mild!!)

- Inference is wrong even under mild violations: we'll tend to reject the null hypothesis of no significance too often when it is true (size is wrong).

- If B is small relative to $\Pi$: the bias will be small.

- If we knew B or (it could be consistently estimated), we could correct the distribution and use it to obtain valid inference.

- But B cannot be consistently estimated!

- An alternative approach: Conley et al. (2012)

## Plausibly exogeneous, Conley et al. (2012)

- Main idea: Relax the exclusion restriction by adopting a Bayesian approach

- Set up

$$y = X\beta + Z\gamma + \epsilon; \quad (18)$$
$$X = Z\Pi + v; \quad (19)$$
$$corr(\epsilon, v) \neq 0 \quad (20)$$

where X can contain various endogenous regressors (s) and Z contains r instruments (r$\geq s$).

# Plausibly exogeneous, II

- Bayesian approach: incorporate beliefs about $\gamma$

- The IV exclusion restriction is equivalent to the dogmatic belief that $\gamma = 0$.

- New belief: $\gamma$ is close to zero, but maybe not identical to zero.

- To implement this technique in STATA: ssc install plausexog

## Plausibly exogeneous, III

- 4 inference strategies that use prior information about $\gamma$

- 1. The research specifies the support of $\gamma$.

- 2 and 3: The research specifies some prior information about the distribution of $\gamma$

- and 4: Full Bayesian approach, with priors over all parameters.

- Their strategy allows to compute confidence intervals for $\beta$ conditional on any potential value of $\gamma$

- This allows to evaluate the robustness to deviations from the exclusion restriction.

## Example (from Conley et al.,)

- Price elasticity of Demand for margarine.

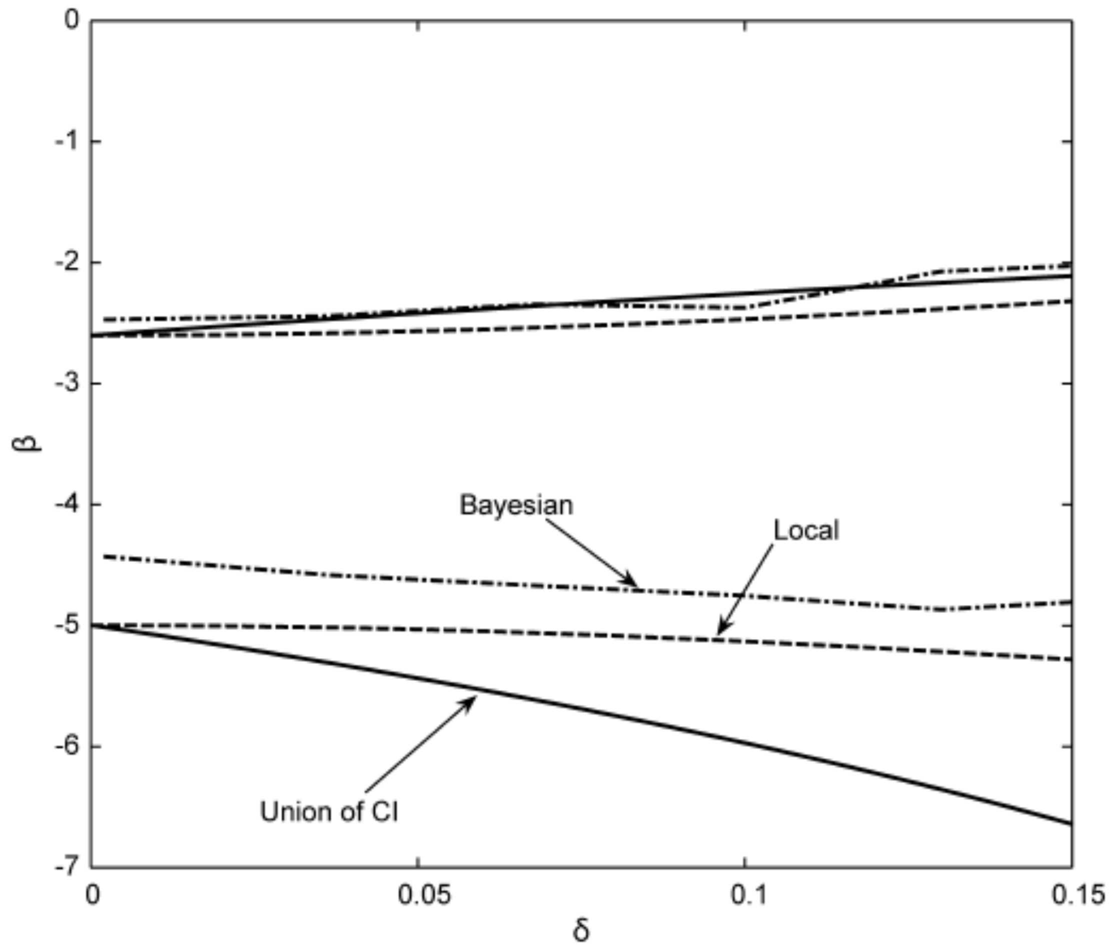
$$log(share) = \beta log(retailprice) + controls + v_t$$

- Instrument: log wholesale prices. Why? they should primarily vary in response to cost shocks and be less sensitive to retail demand shocks than retail prices.

- The following picture shows that there can be substantial violations of the exclusion restriction and the estimates won't vary much.

- $\delta$ measures how large is the deviation from the exclusion restriction (i.e., $\delta = 0 \longrightarrow$ exclusion restriction holds). Its precise definition varies depending on the method employed.

FIGURE 4.—95% INTERVAL ESTIMATES FROM MARGARINE EXAMPLE WITH $\gamma|\beta$ PRIOR




This figure presents 95% confidence intervals for the price elasticity of the demand for margarine across various prior settings. The definition of $\delta$ differs between the different intervals. The "Union of CI" intervals impose only the prior information that the support of $\gamma$ is $[-2\delta|\beta|, 2\delta|\beta|]$. The "Local" and "Bayesian" intervals impose the prior that $\gamma \sim N(0, \delta^2\beta^2)$.

A qualitative conclusion from figure 4 that is common across methods is that there can be a substantial violation of the exclusion restriction without a major change in the demand elasticity estimates. Inferences change little for a range of direct wholesale price effects up to 10% of the size of the retail price effect. Take, for example, the local-to-0 estimates, at $\delta = 0$ the 95% confidence interval is $(-5, -2.5)$ and at $\delta = 10\%$ the 95% confidence interval is $(-5.5, -2.3)$. Put on a standard markup basis using the inverse elasticity rule, the corresponding mark-up intervals are [20% to 40%] and [18% to 44%]. For many, if not all, purposes, this is a small change in the implied mark-ups.

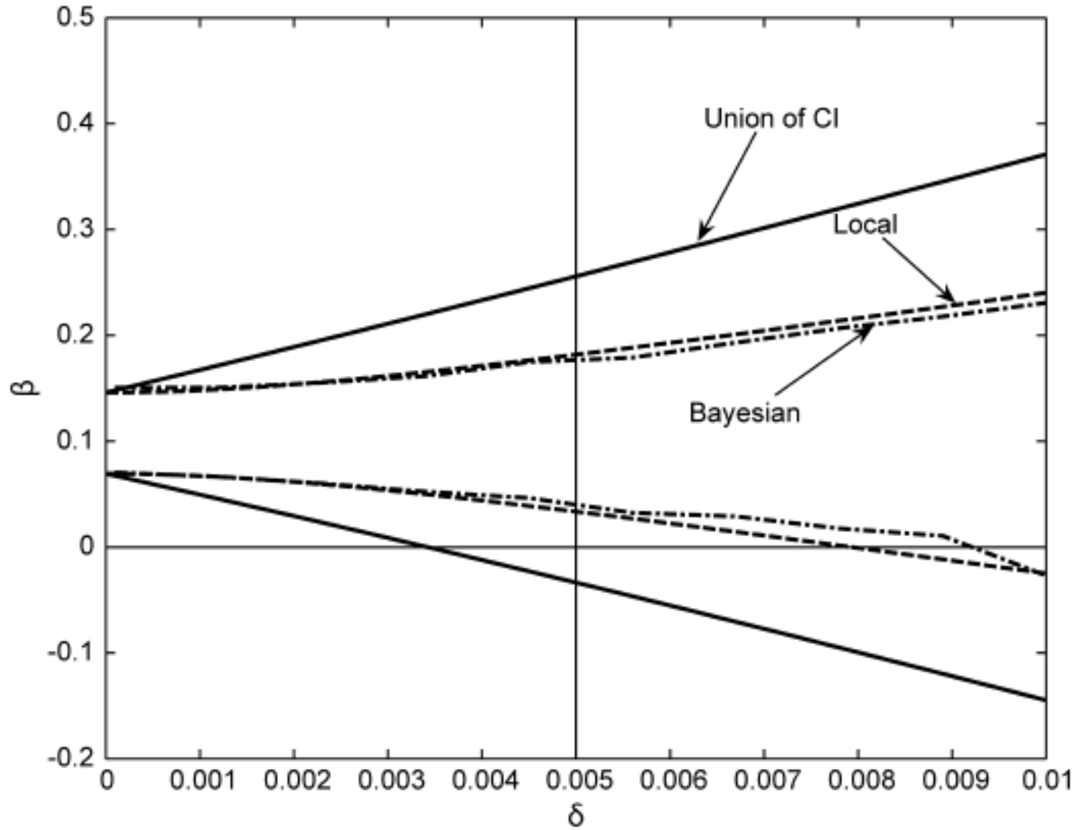
# Example 2: Returns to schooling

- Data from Angrist and Krueger (1991), 300.000+ observations

$$log(wage_i) = \beta_1 education_i + controls + u_i$$

- Conley use some of Bound's arguments/calculations to set up priors for the parameters.

FIGURE 5.—95% INTERVAL ESTIMATES FROM RETURNS-TO-SCHOOLING EXAMPLE




This figure presents 95% confidence intervals for the returns to schooling across various prior settings. The definition of δ differs between the different intervals. The "Union of CI" intervals impose only the prior information that the γ takes on values within the cube $[-2\delta, 2\delta]^3$. The "Local" and "Bayesian" intervals impose the prior that $\gamma \sim N(0, \delta^2 I_3)$ where $I_3$ is a $3 \times 3$ identity matrix.

The intervals in figure 5 suggest that the data are essentially uninformative about the returns to schooling under priors consistent with the evidence in Bound et al. (1995). Using the Bound et al. (1995) calculations as an upper bound on the magnitude of $\gamma$ would require us to focus attention in a $\delta$ range near .005. At $\delta = .005$, the local-to-0 95% confidence interval for $\beta$ is [3.4% to 18.3%], which we consider uninformative about the returns to years of school. In order for these confidence intervals to be informative in our judgment, prior beliefs regarding $\gamma$ must be much more concentrated near 0. For example, using the support-restriction-only intervals, one would need to be sure that the magnitude of $\gamma$ was less than .002 to obtain a confidence interval for $\beta$ that excluded 5%.

- Last remark:

The stronger the instruments, the more robust your estimates will be to deviations from the exclusion restriction

- Since the strength of the instruments is something that can be evaluated (as opposed to the "exogeneity") is important to have instruments as strong as possible!

## What you should do in practice?

[Advice from *Mostly Harmless...*]

- Report the first stage and think about whether it makes sense. Are the magnitude and sign as you would expect?
- Report the F-statistics on the excluded instruments. The bigger it is the better. Use a proper test
- Pick your best single instrument and report just-identified estimates using this one only. Just-identified IV is approximately median-unbiased
- Check over-identified 2SLS estimates with LIML. If the LIML estimates are very different, or standard errors are much bigger, worry.

  - See also Murray (2006) for more practical advice!