ECON 6200: Econometrics II

Instrumental Variables

Instrumental variable techniques are of overwhelming importance in economics.

They are also one a major contribution of econometrics to empirical methods more generally.

They are actively used in causal inference across disciplines and specifically in biostatistics ("Mendelian randomization").

Their appeal is that they allow for causally interpretable estimates if we think that

- the Linear Model (or, later, generalizations thereof) is structural and β therefore has causal interpretation...
- but ε correlates with X, e.g. because ε absorbs relevant omitted variables, due to simultaneous causation, etc.

Of course, there is no free lunch:
This remarkable result requires restrictive assumptions.

Instrumental Variables on One Slide

For simplicity only, consider simple linear regression:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

but where we cannot assert $\mathbb{E} X \varepsilon = 0$.

Again, the interpretation is that:

- β_1 has a causal interpretation, i.e. there is a true linear model, however...
- we do not observe all relevant covariates and therefore estimating the above by OLS would incur omitted variable bias.

Instrumental Variables on One Slide

For simplicity only, consider simple linear regression:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

but where we cannot assert $\mathbb{E} X\varepsilon = 0$.

Again, the interpretation is that:

- β_1 has a causal interpretation, i.e. there is a true linear model, however...
- we do not observe all relevant covariates and therefore estimating the above by OLS would incur omitted variable bias.

Now suppose we also observe a r.v. Z with the following properties:

$$\begin{aligned} \text{cov}(Z, X) &\neq 0 &\qquad \leftarrow \text{ relevance} \\ \text{cov}(Z, \varepsilon) &= 0 &\qquad \leftarrow \text{ validity} \end{aligned}$$

Instrumental Variables on One Slide

For simplicity only, consider simple linear regression:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

but where we cannot assert $\mathbb{E} X\varepsilon = 0$.

Again, the interpretation is that:

- β_1 has a causal interpretation, i.e. there is a true linear model, however...
- we do not observe all relevant covariates and therefore estimating the above by OLS would incur omitted variable bias.

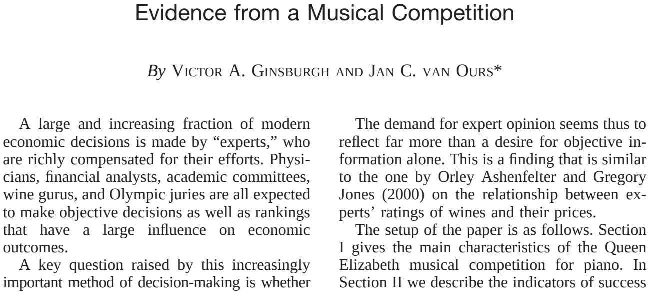
Now suppose we also observe a r.v. Z with the following properties:

$$\begin{aligned} \operatorname{cov}(Z, X) &\neq 0 \qquad \leftarrow \text{ relevance} \\ \operatorname{cov}(Z, \varepsilon) &= 0 \qquad \leftarrow \text{ validity} \end{aligned}$$

$$\implies \frac{\operatorname{cov}(Z, Y)}{\operatorname{cov}(Z, X)} = \frac{\beta_1 \operatorname{cov}(Z, X) + \operatorname{cov}(Z, \varepsilon)}{\operatorname{cov}(Z, X)} = \beta_1 .$$

Instrumental Variables

Empirical Example



Evidence from a Musical Competition

By Victor A. Ginsburgh and Jan C. van Ours*

A large and increasing fraction of modern economic decisions is made by "experts," who are richly compensated for their efforts. Physicians, financial analysts, academic committees, wine gurus, and Olympic juries are all expected to make objective decisions as well as rankings that have a large influence on economic outcomes.

A key question raised by this increasingly important method of decision-making is whether

The demand for expert opinion seems thus to reflect far more than a desire for objective information alone. This is a finding that is similar to the one by Orley Ashenfelter and Gregory Jones (2000) on the relationship between experts' ratings of wines and their prices.

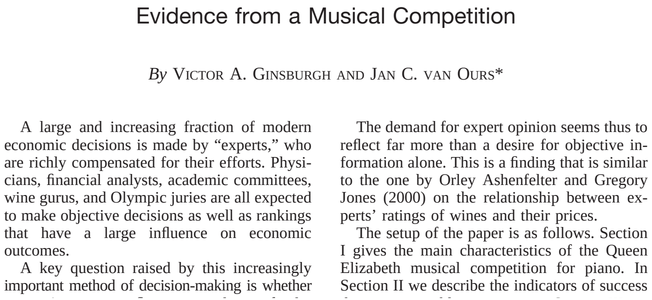
The setup of the paper is as follows. Section I gives the main characteristics of the Queen Elizabeth musical competition for piano. In Section II we describe the indicators of success

In this example (*American Economic Review*, 2003):

- Y=career success of a classical musician,
- X=placement in a prestigious competition for young musicians,
- Z=order of appearance at said competition.

Instrumental Variables

Empirical Example



Evidence from a Musical Competition

By Victor A. Ginsburgh and Jan C. van Ours*

A large and increasing fraction of modern economic decisions is made by “experts,” who are richly compensated for their efforts. Physicians, financial analysts, academic committees, wine gurus, and Olympic juries are all expected to make objective decisions as well as rankings that have a large influence on economic outcomes.

A key question raised by this increasingly important method of decision-making is whether

The demand for expert opinion seems thus to reflect far more than a desire for objective information alone. This is a finding that is similar to the one by Orley Ashenfelter and Gregory Jones (2000) on the relationship between experts’ ratings of wines and their prices.

The setup of the paper is as follows. Section I gives the main characteristics of the Queen Elizabeth musical competition for piano. In Section II we describe the indicators of success

The effect of X on Y is interesting but a musician’s talent will impact both.

However, appearing late in the competition (verifiably) predicts success.

Yet order of appearance is randomized!

It therefore serves as instrumental variable or just "instrument."

Instrumental Variables

More on Relevance

Again, Relevance requires that $\text{cov}(Z, X) \neq 0$:
"The instrumental variable is correlated with the endogenous regressor."

This ensures that we have instrument-induced variation to play with.

If it were not required, the following pathological conclusion would arise:

Instrumental Variables

More on Relevance

Again, Relevance requires that $\text{cov}(Z, X) \neq 0$:
"The instrumental variable is correlated with the endogenous regressor."

This ensures that we have instrument-induced variation to play with.

If it were not required, the following pathological conclusion would arise:
I could just have my Research Assistant flip coins all day and use that as instrument.

Instrumental Variables

More on Relevance

Again, Relevance requires that $\text{cov}(Z, X) \neq 0$:
"The instrumental variable is correlated with the endogenous regressor."

This ensures that we have instrument-induced variation to play with.

If it were not required, the following pathological conclusion would arise:
I could just have my Research Assistant flip coins all day and use that as instrument.

Relevance is **testable**.

- The covariance is consistently estimated by its sample analog.
- Indeed, it is standard practice to report the F-statistic from a "first-stage regression" of X on Z.

Instrumental Variables

More on Validity

Validity requires that $\text{cov}(Z, \varepsilon) = 0$:
"The instrumental variable is not correlated with unobservables."

This ensures that the instrument-induced variation is exogenous.

If it were not required, the following pathological conclusion would arise:

Instrumental Variables

More on Validity

Validity requires that $\text{cov}(Z, \varepsilon) = 0$:
"The instrumental variable is not correlated with unobservables."

This ensures that the instrument-induced variation is exogenous.

If it were not required, the following pathological conclusion would arise:
I could just use X as instrument for itself.

(Indeed, OLS *is* the special case of IV where X is its own instrument.)

Instrumental Variables

More on Validity

Validity requires that $\text{cov}(Z, \varepsilon) = 0$:
"The instrumental variable is not correlated with unobservables."

This ensures that the instrument-induced variation is exogenous.

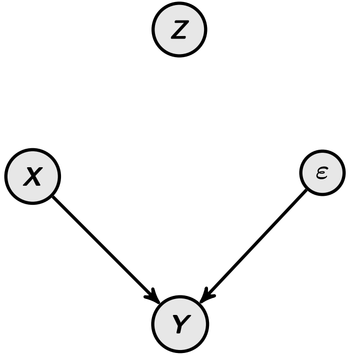
If it were not required, the following pathological conclusion would arise:
I could just use X as instrument for itself.

(Indeed, OLS *is* the special case of IV where X is its own instrument.)

Validity is **not testable**.
(This assessment will change when we have more instruments than regressors.)

Instrumental Variables

Illustration using Directed Acyclic Graphs

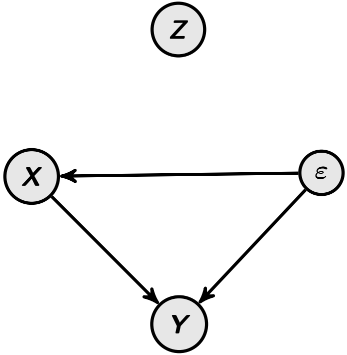


This is a Directed Acyclic Graph (DAG) illustrating OLS. Arrows read as "... causes ...," where causality is defined through conditional independence.

All is good here. The random variable Z plays no role yet.

Instrumental Variables

Illustration using Directed Acyclic Graphs

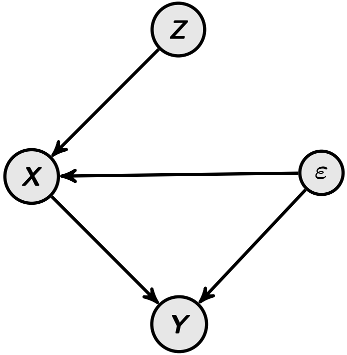


This is linear regression, except that X is **endogenous**.

The observable association between X and Y is **confounded** by (something hiding in) ε.

Instrumental Variables

Illustration using Directed Acyclic Graphs



This is the usual DAG representation of Instrumental Variables.

(If you want to dig deeper into modeling causality like this, look up **graphical models** or the work of computer scientist Judea Pearl.)

Instrumental Variables

Aside: Instrumental Variables in medical statistics

Mendelian Randomization is an Instrumental Variables technique.

The idea is that, in our terminology, genetic variation may be exogenous.

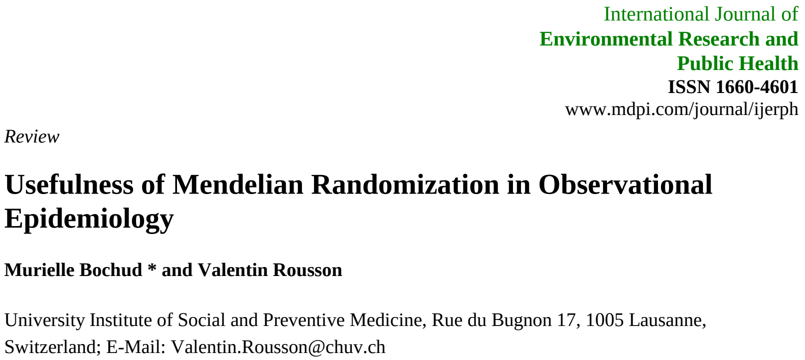
To get the idea, think:

- $Y =$ coronary disease,
- $X =$ HDL cholesterol level,
- $\varepsilon =$ confounders
  (there will be many: diet and lifestyle heavily affect HDL cholesterol level),
- $Z =$ genetic variant that increases HDL cholesterol level.

The bottleneck assumption is that the genetic variant *only* affects HDL cholesterol level.

Instrumental Variables

Instrumental Variables in medical statistics



International Journal of
**Environmental Research and
Public Health**
ISSN 1660-4601
www.mdpi.com/journal/ijerph

Review

Usefulness of Mendelian Randomization in Observational Epidemiology

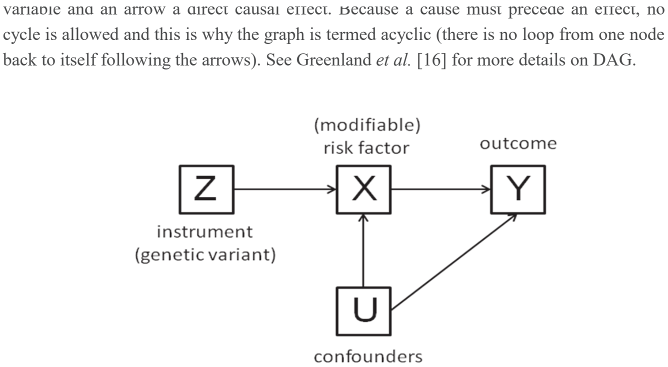
Murielle Bochud * and Valentin Rousson

University Institute of Social and Preventive Medicine, Rue du Bugnon 17, 1005 Lausanne,
Switzerland; E-Mail: Valentin.Rousson@chuv.ch

In this survey on Mendelian Randomization...

Instrumental Variables

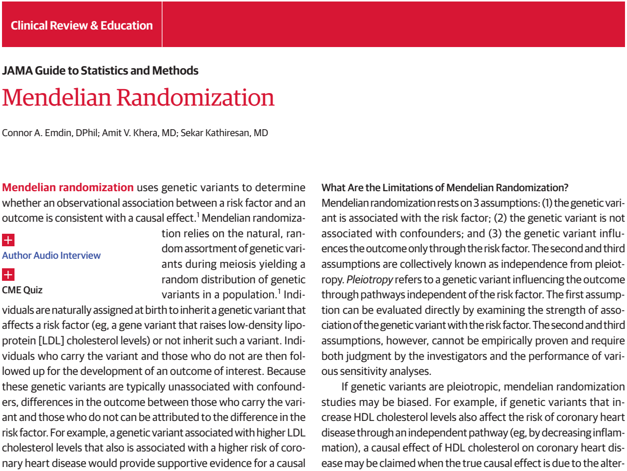
Instrumental Variables in medical statistics



variable and an arrow a direct causal effect. Because a cause must precede an effect, no cycle is allowed and this is why the graph is termed acyclic (there is no loop from one node back to itself following the arrows). See Greenland *et al.* [16] for more details on DAG.

...we find this visualization.

Instrumental Variables in medical statistics



Clinical Review & Education

JAMA Guide to Statistics and Methods

Mendelian Randomization

Connor A. Emdin, DPhil; Amit V. Khera, MD; Sekar Kathiresan, MD

Mendelian randomization uses genetic variants to determine whether an observational association between a risk factor and an outcome is consistent with a causal effect.[1] Mendelian randomization relies on the natural, random assortment of genetic variants during meiosis yielding a random distribution of genetic variants in a population.[1] Individuals are naturally assigned at birth to inherit a genetic variant that affects a risk factor (eg, a gene variant that raises low-density lipoprotein [LDL] cholesterol levels) or not inherit such a variant. Individuals who carry the variant and those who do not are then followed up for the development of an outcome of interest. Because these genetic variants are typically unassociated with confounders, differences in the outcome between those who carry the variant and those who do not can be attributed to the difference in the risk factor. For example, a genetic variant associated with higher LDL cholesterol levels that also is associated with a higher risk of coronary heart disease would provide supportive evidence for a causal

Author Audio Interview

CME Quiz

What Are the Limitations of Mendelian Randomization?

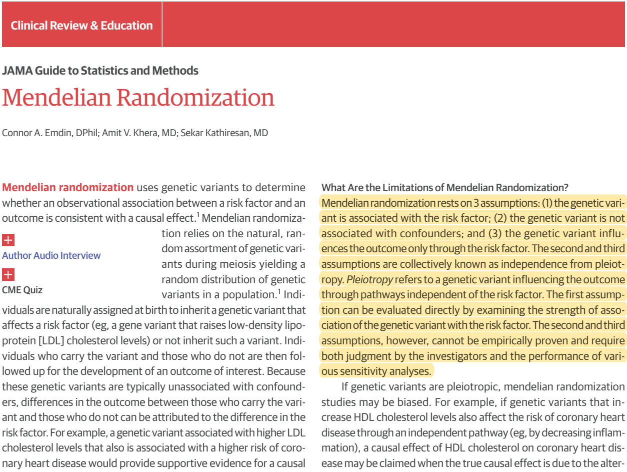
Mendelian randomization rests on 3 assumptions: (1) the genetic variant is associated with the risk factor; (2) the genetic variant is not associated with confounders; and (3) the genetic variant influences the outcome only through the risk factor. The second and third assumptions are collectively known as independence from pleiotropy. *Pleiotropy* refers to a genetic variant influencing the outcome through pathways independent of the risk factor. The first assumption can be evaluated directly by examining the strength of association of the genetic variant with the risk factor. The second and third assumptions, however, cannot be empirically proven and require both judgment by the investigators and the performance of various sensitivity analyses.

If genetic variants are pleiotropic, mendelian randomization studies may be biased. For example, if genetic variants that increase HDL cholesterol levels also affect the risk of coronary heart disease through an independent pathway (eg, by decreasing inflammation), a causal effect of HDL cholesterol on coronary heart disease may be claimed when the true causal effect is due to the alter-

In this recent exposition (*Journal of the American Medical Association* 2019; also the source of my example),...

Instrumental Variables

Instrumental Variables in medical statistics



Clinical Review & Education

JAMA Guide to Statistics and Methods

Mendelian Randomization

Connor A. Emdin, DPhil; Amit V. Khera, MD; Sekar Kathiresan, MD

Mendelian randomization uses genetic variants to determine whether an observational association between a risk factor and an outcome is consistent with a causal effect.[1] Mendelian randomization relies on the natural, random assortment of genetic variants during meiosis yielding a random distribution of genetic variants in a population.[1] Individuals are naturally assigned at birth to inherit a genetic variant that affects a risk factor (eg, a gene variant that raises low-density lipoprotein [LDL] cholesterol levels) or not inherit such a variant. Individuals who carry the variant and those who do not are then followed up for the development of an outcome of interest. Because these genetic variants are typically unassociated with confounders, differences in the outcome between those who carry the variant and those who do not can be attributed to the difference in the risk factor. For example, a genetic variant associated with higher LDL cholesterol levels that also is associated with a higher risk of coronary heart disease would provide supportive evidence for a causal

+ Author Audio Interview

+ CME Quiz

What Are the Limitations of Mendelian Randomization?

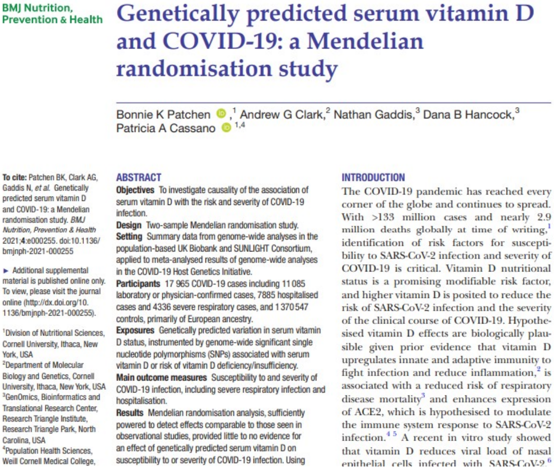
Mendelian randomization rests on 3 assumptions: (1) the genetic variant is associated with the risk factor; (2) the genetic variant is not associated with confounders; and (3) the genetic variant influences the outcome only through the risk factor. The second and third assumptions are collectively known as independence from pleiotropy. *Pleiotropy* refers to a genetic variant influencing the outcome through pathways independent of the risk factor. The first assumption can be evaluated directly by examining the strength of association of the genetic variant with the risk factor. The second and third assumptions, however, cannot be empirically proven and require both judgment by the investigators and the performance of various sensitivity analyses.

If genetic variants are pleiotropic, mendelian randomization studies may be biased. For example, if genetic variants that increase HDL cholesterol levels also affect the risk of coronary heart disease through an independent pathway (eg, by decreasing inflammation), a causal effect of HDL cholesterol on coronary heart disease may be claimed when the true causal effect is due to the alter-

...we find a discussion that should read familiar.

Instrumental Variables

Instrumental Variables in medical statistics



BMJ Nutrition, Prevention & Health

Genetically predicted serum vitamin D and COVID-19: a Mendelian randomisation study

Bonnie K Patchen,[1] Andrew G Clark,[2] Nathan Gaddis,[3] Dana B Hancock,[3] Patricia A Cassano[1,4]

To cite: Patchen BK, Clark AG, Gaddis N, *et al.* Genetically predicted serum vitamin D and COVID-19: a Mendelian randomisation study. *BMJ Nutrition, Prevention & Health* 2021;**4**:e000255. doi:10.1136/bmjnph-2021-000255

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjnph-2021-000255).

[1]Division of Nutritional Sciences, Cornell University, Ithaca, New York, USA
[2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA
[3]GenOmics, Bioinformatics and Translational Research Center, Research Triangle Institute, Research Triangle Park, North Carolina, USA
[4]Population Health Sciences, Weill Cornell Medical College,

ABSTRACT

Objectives To investigate causality of the association of serum vitamin D with the risk and severity of COVID-19 infection.
Design Two-sample Mendelian randomisation study.
Setting Summary data from genome-wide analyses in the population-based UK Biobank and SUNLIGHT Consortium, applied to meta-analysed results of genome-wide analyses in the COVID-19 Host Genetics Initiative.
Participants 17 965 COVID-19 cases including 11 085 laboratory or physician-confirmed cases, 7885 hospitalised cases and 4336 severe respiratory cases, and 1 370 547 controls, primarily of European ancestry.
Exposures Genetically predicted variation in serum vitamin D status, instrumented by genome-wide significant single nucleotide polymorphisms (SNPs) associated with serum vitamin D or risk of vitamin D deficiency/insufficiency.
Main outcome measures Susceptibility to and severity of COVID-19 infection, including severe respiratory infection and hospitalisation.
Results Mendelian randomisation analysis, sufficiently powered to detect effects comparable to those seen in observational studies, provided little to no evidence for an effect of genetically predicted serum vitamin D on susceptibility to or severity of COVID-19 infection. Using

INTRODUCTION

The COVID-19 pandemic has reached every corner of the globe and continues to spread. With >133 million cases and nearly 2.9 million deaths globally at time of writing,[1] identification of risk factors for susceptibility to SARS-CoV-2 infection and severity of COVID-19 is critical. Vitamin D nutritional status is a promising modifiable risk factor, and higher vitamin D is posited to reduce the risk of SARS-CoV-2 infection and the severity of the clinical course of COVID-19. Hypothesised vitamin D effects are biologically plausible given prior evidence that vitamin D upregulates innate and adaptive immunity to fight infection and reduce inflammation,[2] is associated with a reduced risk of respiratory disease mortality[3] and enhances expression of ACE2, which is hypothesised to modulate the immune system response to SARS-CoV-2 infection.[4 5] A recent in vitro study showed that vitamin D reduces viral load of nasal epithelial cells infected with SARS-CoV-2.[6]

A recent example made at Cornell. (The result is negative.)

Instrumental Variables

Some more examples

Before we embark on a technical analysis of IV, some more examples.

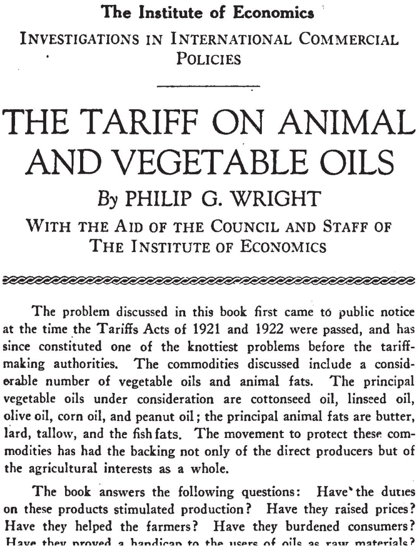
The following is a "greatest hits" of famous examples presented extremely briefly.

"Famous" need not imply good!

Several have been criticized. Think about them for yourself.

Instrumental Variables

How it all began



The Institute of Economics

INVESTIGATIONS IN INTERNATIONAL COMMERCIAL POLICIES

THE TARIFF ON ANIMAL AND VEGETABLE OILS

By PHILIP G. WRIGHT

WITH THE AID OF THE COUNCIL AND STAFF OF THE INSTITUTE OF ECONOMICS

The problem discussed in this book first came to public notice at the time the Tariffs Acts of 1921 and 1922 were passed, and has since constituted one of the knottiest problems before the tariff-making authorities. The commodities discussed include a considerable number of vegetable oils and animal fats. The principal vegetable oils under consideration are cottonseed oil, linseed oil, olive oil, corn oil, and peanut oil; the principal animal fats are butter, lard, tallow, and the fish fats. The movement to protect these commodities has had the backing not only of the direct producers but of the agricultural interests as a whole.

The book answers the following questions: Have the duties on these products stimulated production? Have they raised prices? Have they helped the farmers? Have they burdened consumers? Have they proved a handicap to the users of oils as raw materials?

Instrumental Variables

How it all began

Wright (1928) wants to estimate supply and demand elasticity for vegetable oils.

He recognizes that observed quantities and prices do not trace either curve but are an equilibrium phenomenon.

He identifies numerous "demand shifters" (price changes of substitute goods) and "supply shifters" (weather) and uses them as instruments.

He averages the estimators – not something we would do today.

Instrumental Variables

Quarter of Birth as Instrument for Education

(Angrist & Krueger, *Quarterly Journal of Economics*, 1991)

- **Aim of inference:**
  Returns to schooling.
- **The problem:**
  Selection into schooling.
- **What is the instrument?**
  Quarter of birth.
- **Why might I believe the instrument?**
  Compulsory schooling may effectively last one year longer depending on when in a year one is born.

Instrumental Variables

Child Gender as Instrument for Fertility

(Angrist & Evans, *American Economic Review*, 1998)

- **Aim of inference:**
  Effect of having kids on female labor force participation.
- **The problem:**
  Decision to have kids is endogenous.
- **What is the instrument?**
  Gender composition of a household's children.
- **Why might I believe the instrument?**
  This is quite clearly random.
  But many parents prefer to have children of both genders, and so probability of a third child is higher if the first two were of same gender.

Instrumental Variables

Proximity to College as Instrument for College Education

(Card, a book chapter, 1995)

- **Aim of inference:**
  Private returns to College education.
- **The problem:**
  College attendance is endogenous.
- **What is the instrument?**
  Growing up near a college.
- **Why might I believe the instrument?**
  Because in the big picture of things, college proximity might rarely be pivotal in mobility decisions.
  (But this example is already less clear-cut than others!)

Instrumental Variables

Analyzing the IV Estimator

We next derive the IV (and Two-Stage Least Squares) estimator as Generalized Method of Moments estimator.

Recall the method of moments: If we assume

$$\begin{aligned} Y &= X'\beta + \varepsilon \\ \mathbb{E}(X\varepsilon) &= 0 \\ \Longrightarrow \mathbb{E}(X(Y - X'\beta) &= 0, \end{aligned}$$

then a natural idea for estimating β is to solve the empirical moment condition

$$\mathbb{E}_n\big(X(Y - X'\hat{\beta})\big) = 0.$$

This is easily seen to yield the OLS estimator.
(The above is the FOC of the least squares minimization problem.)

Instrumental Variables

Analyzing the IV Estimator

Now assume

$$Y = X'\beta + \varepsilon$$

but with

$$\mathbb{E}(X\varepsilon) \neq 0.$$

OLS will be inconsistent for β because it will estimate the projection coefficient

$$b^* = (\mathbb{E}XX')^{-1}\mathbb{E}XY = \beta + (\mathbb{E}XX')^{-1}\mathbb{E}X\varepsilon.$$

Instrumental Variables

Analyzing the IV Estimator

Now assume

$$Y = X'\beta + \varepsilon$$

but with

$$\mathbb{E}(X\varepsilon) \neq 0.$$

OLS will be inconsistent for β because it will estimate the projection coefficient

$$b^* = (\mathbb{E}XX')^{-1}\mathbb{E}XY = \beta + (\mathbb{E}XX')^{-1}\mathbb{E}X\varepsilon.$$

However, we also observe a random k-vector Z with

$$\begin{aligned} \mathbb{E}Z\varepsilon &= 0 \\ \text{rank}(\mathbb{E}ZX') &= k. \end{aligned}$$

Instrumental Variables

Analyzing the IV Estimator

Now assume

$$Y = X'\beta + \varepsilon$$

but with

$$\mathbb{E}(X\varepsilon) \neq 0.$$

OLS will be inconsistent for β because it will estimate the projection coefficient

$$b^* = (\mathbb{E}XX')^{-1}\mathbb{E}XY = \beta + (\mathbb{E}XX')^{-1}\mathbb{E}X\varepsilon.$$

However, we also observe a random k-vector Z with

$$\begin{aligned} \mathbb{E}Z\varepsilon &= 0 \\ \text{rank}(\mathbb{E}ZX') &= k. \end{aligned}$$

Then we have moment condition

$$\begin{aligned} \mathbb{E}(Z(Y - X'\beta)) &= 0 \\ \implies \beta &= (\mathbb{E}ZX')^{-1}\mathbb{E}ZY, \end{aligned}$$

where assumptions ensure that the last expression is well-defined.

Instrumental Variables

The population-level finding

$$\beta = (\mathbb{E}ZX')^{-1}\mathbb{E}ZY$$

inspires estimator

$$\hat{\beta}_{IV} = (\mathbb{E}_n ZX')^{-1}\mathbb{E}_n ZY.$$

Note that

$$\hat{\beta}_{IV} = (\mathbb{E}_n ZX')^{-1}\mathbb{E}_n ZY = (\mathbb{E}_n ZX')^{-1}\mathbb{E}_n Z(X'\beta + \varepsilon) = \beta + \underbrace{(\mathbb{E}_n ZX')^{-1}\mathbb{E}_n Z\varepsilon}_{= \text{estimation error}}.$$

- From here, consistency follows essentially as before, now from $\mathbb{E}_n Z\varepsilon \xrightarrow{p} 0$.
- However, we cannot claim unbiasedness, not even if we were to assume $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{Z}) = \mathbf{0}$. Why?
- We will defer full development of asymptotic theory until after generalizing this estimator.

Instrumental Variables

The population-level finding

$$\beta = (\mathbb{E} ZX')^{-1} \mathbb{E} ZY$$

inspires estimator

$$\hat{\beta}_{IV} = (\mathbb{E}_n ZX')^{-1} \mathbb{E}_n ZY.$$

Note that

$$\hat{\beta}_{IV} = (\mathbb{E}_n ZX')^{-1} \mathbb{E}_n ZY = (\mathbb{E}_n ZX')^{-1} \mathbb{E}_n Z(X'\beta + \varepsilon) = \beta + \underbrace{(\mathbb{E}_n ZX')^{-1} \mathbb{E}_n Z\varepsilon}_{= \text{estimation error}}.$$

- Some components of Z may also appear in X. This is the case when some but not all covariates are endogenous. Exogenous covariates effectively act as their own instruments.
- By setting $Z = X$, we can consider OLS the special case of all covariates instrumenting for themselves.

Instrumental Variables

Some Other Representations of the Estimator

In the simple linear model, i.e. when X and Z are scalars, the IV slope estimator can be expressed as

$$\frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})}.$$

Instrumental Variables

Some Other Representations of the Estimator

In the simple linear model, i.e. when X and Z are scalars, the IV slope estimator can be expressed as

$$\frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})}.$$

The data matrix expression for the estimator is $\hat{\beta}_{IV} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.

Instrumental Variables

Some Other Representations of the Estimator

In the simple linear model, i.e. when X and Z are scalars, the IV slope estimator can be expressed as

$$\frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})}.$$

The data matrix expression for the estimator is $\hat{\beta}_{IV} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.

For an application of the latter, consider first regressing X on Z and then Y on $\hat{X}$. Written in data matrix notation, the "second stage" estimator is

$$\begin{aligned}
\tilde{\beta} &= (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y} \\
&= \left((\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}(\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})'\boldsymbol{Y} \\
&= \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\
&= \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\
&= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\
&= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \hat{\beta}_{IV}.
\end{aligned}$$

Instrumental Variables

IV as Two-Stage Regression

Let us think more about the observation that

$$\tilde{\beta} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \hat{\beta}_{IV}.$$

- The IV estimator can be thought of as a two-stage regression whereby we first project X on Z and then project Y on the fitted values $\hat{X}$.
- This gives an intuition: We exploit only that variation in X that is "due to" (in the sense of correlation, causality is unclear here!) Z.
- The interpretation can also be related to the DAG representation.
- It is why the regression of X on Z is called "first-stage regression." It is usually reported. To suggest relevance of Z, it should be highly significant. A rule of thumb is $F = 10$ for the overall first-stage regression.
- Note: The first stage is purely correlational at this point. This is a point to which we may return, e.g. when looking at optimal instruments or "machine learning" the first stage.

Instrumental Variables

IV as Two-Stage Regression

Let us think more about the observation that

$$\tilde{\beta} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \hat{\beta}_{IV}.$$

- The idea does not require X and Z to be of same length.
- Thus, we can do IV with more instruments than regressors!
- Caveat: If $\boldsymbol{Z}'\boldsymbol{X}$ is not square (but assuming it has maximal rank), the algebraic derivation gets stuck at

$$\begin{aligned}
\tilde{\beta} &= (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'Y \\
&= \left((\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}(\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})'\boldsymbol{Y} \\
&= \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\
&= \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \equiv \hat{\beta}_{TSLS}.
\end{aligned}$$

- This is the Two-Stage Least Squares estimator.

Instrumental Variables

IV versus TSLS

A major difference between the estimators is that, for IV,

$$\boldsymbol{Z}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = \boldsymbol{0}$$

by construction as before.
Can this also be true with TSLS?

Instrumental Variables

IV versus TSLS

A major difference between the estimators is that, for IV,

$$\boldsymbol{Z}'\hat{\varepsilon} = \boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = \mathbf{0}$$

by construction as before.
Can this also be true with TSLS?

No! With $\ell > k$ instruments, the above are ℓ linear equations in k unknowns.

This has a number of major implications:

- The estimator cannot set $\boldsymbol{Z}'\hat{\varepsilon} = \mathbf{0}$.
- Can show: The estimator minimizes (in b) some norm $\|\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}b)\|$.
  But is this the best norm?
- Validity of instruments becomes testable:
  With large samples, should have $\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon} \approx \mathbf{0}$.

Instrumental Variables

IV versus TSLS

A major difference between the estimators is that, for IV,

$$\boldsymbol{Z}'\hat{\varepsilon} = \boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = \mathbf{0}$$

by construction as before.
Can this also be true with TSLS?

No! With $\ell > k$ instruments, the above are ℓ linear equations in k unknowns.

This has a number of major implications:

- The estimator cannot set $\boldsymbol{Z}'\hat{\varepsilon} = \mathbf{0}$.
- Can show: The estimator minimizes (in b) some norm $\|\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}b)\|$.
  But is this the best norm?
- Validity of instruments becomes testable:
  With large samples, should have $\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon} \approx \mathbf{0}$.

These considerations lead us from TSLS to the Generalized Method of Moments.
We will soon develop this method in detail.
Detailed asymptotic characterization of TSLS will be a corollary.

Asymptotics of Simple IV

We briefly develop the asymptotics of simple instrumental variables, i.e.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where (Y, X, z) are scalars, and furthermore assume homoskedasticity. Reasons:

- The asymptotic variance is instructive,
- This allows us to formally talk about weak instruments.

Instrumental Variables

Asymptotics of Simple IV

We briefly develop the asymptotics of simple instrumental variables, i.e.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where (Y, X, z) are scalars, and furthermore assume homoskedasticity. Reasons:

- The asymptotic variance is instructive,
- This allows us to formally talk about weak instruments.

Plugging in from previous algebra,

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_1 - \beta_1) &= \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)} \\
&= \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \mathbb{E}Z\right)\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \mathbb{E}Z\right)\left(X_i - \mathbb{E}X\right)} + o_P(1) \\
&\overset{d}{\to} \frac{N(0, \sigma_z^2\sigma^2)}{\rho\sigma_z\sigma_x} = N\left(0, \frac{\sigma^2}{\rho^2\sigma_x^2}\right).
\end{aligned}$$

Instrumental Variables

Compare

$$\begin{aligned}\sqrt{n}(\hat{\beta}_1^{IV} - \beta_1) &\xrightarrow{d} N\left(0, \frac{\sigma^2}{\rho^2\sigma_x^2}\right)\\ \sqrt{n}(\hat{\beta}_1^{OLS} - b_1^*) &\xrightarrow{d} N\left(0, \frac{\sigma^2}{\sigma_x^2}\right)\end{aligned}$$

Instrumental Variables

Compare

$$\begin{aligned}\sqrt{n}(\hat{\beta}_1^{IV} - \beta_1) &\xrightarrow{d} N\left(0, \frac{\sigma^2}{\rho^2\sigma_x^2}\right)\\ \sqrt{n}(\hat{\beta}_1^{OLS} - b_1^*) &\xrightarrow{d} N\left(0, \frac{\sigma^2}{\sigma_x^2}\right)\end{aligned}$$

- The result has an easy intuition:
  The IV asymptotic variance has the "first-stage population explained variation" in the denominator.
- However, note that this is not a result about finite-sample variance!
- In fact, the finite-sample variance of $\hat{\beta}_1^{IV}$ does not exist.

Instrumental Variables

Compare

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_1^{IV} - \beta_1) &\xrightarrow{d} N\left(0, \frac{\sigma^2}{\rho^2\sigma_x^2}\right) \\
\sqrt{n}(\hat{\beta}_1^{OLS} - b_1^*) &\xrightarrow{d} N\left(0, \frac{\sigma^2}{\sigma_x^2}\right)
\end{aligned}$$

- A corollary of the result:
  The IV estimator has higher asymptotic variance.
- The trade-off is against bias.
- If $\rho^2 = 1$, the asymptotic variances coincide.
  (Indeed, the estimators algebraically coincide in this case.)
- The asymptotic variance diverges toward ∞ as $\rho \to 0$.
  This suggests that the result is not uniform as the instrument becomes weak.
  We will next elaborate this.

Instrumental Variables

Weak Instruments

To formally model a weak instrument, set $\rho_n = \rho/\sqrt{n}$.

This is a device:

- Asymptotic approximation is powerful as it allows us to invoke CLT et al.
- But in pointwise perspective, it trivializes the current problem as previous asymptotics go through for any $\rho \neq 0$.
- "Parameter drift" allows us to invoke some asymptotic approximations without "approximating away" the problem.
- Compare Pitman drift for analyzing local power of hypothesis tests.
- The idea is *not* that parameters actually change with n.
  However, the device reinforces that in practice, whether your instrument is weak depends on ρ in relation to n.
  See also Art Goldberger on "micronumerosity."

Instrumental Variables

We develop this for scalar (X, Z). See book for general version.

Write out the first- and second stage regressions:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon \\ X &= \gamma_0 + \frac{\gamma_1}{\sqrt{n}} Z + \eta. \end{aligned}$$

Instrumental Variables

We develop this for scalar (X, Z). See book for general version.

Write out the first- and second stage regressions:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon \\ X &= \gamma_0 + \tfrac{\gamma_1}{\sqrt{n}} Z + \eta. \end{aligned}$$

Then

$$\hat{\beta}_1 - \beta_1 = \frac{\sqrt{n}\mathbb{E}_n(Z - \overline{Z})\varepsilon}{\sqrt{n}\mathbb{E}_n(Z - \overline{Z})(X - \overline{X})},$$

but

$$\begin{aligned} \sqrt{n}\mathbb{E}_n(Z - \overline{Z})(X - \overline{X}) &= \sqrt{n}\mathbb{E}_n(Z - \overline{Z})(\gamma_0 + \gamma_1 / \sqrt{n} Z + \eta) \\ &= \underbrace{\gamma_1 \mathbb{E}_n(Z - \overline{Z})Z}_{\xrightarrow{p} \gamma_1 \sigma_z^2} + \sqrt{n}\mathbb{E}_n(Z - \overline{Z})\eta. \end{aligned}$$

Instrumental Variables

We develop this for scalar (X, Z). See book for general version.

Write out the first- and second stage regressions:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon \\ X &= \gamma_0 + \tfrac{\gamma_1}{\sqrt{n}} Z + \eta. \end{aligned}$$

Assuming CLT applies, conclude that

$$\begin{aligned} \hat{\beta}_1 - \beta_1 &\overset{d}{\to} \frac{a}{\gamma_1 \sigma_z^2 + b}, \\ \begin{pmatrix} a \\ b \end{pmatrix} &\sim N(0, \Omega), \end{aligned}$$

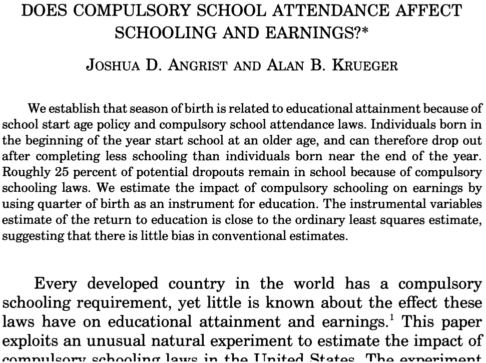
where Ω is the variance-covariance matrix of $(Z\varepsilon, Z\eta)$.

The estimator is inconsistent; indeed, it converges to a distribution!

Observe that the extent of this problem scales inversely with $|\gamma_1|$.

The problem goes away as $|\gamma_1| \to \infty$. This mirrors the breakdown of consistency as the unscaled first-stage slope coefficient vanishes.

Instrumental Variables



DOES COMPULSORY SCHOOL ATTENDANCE AFFECT SCHOOLING AND EARNINGS?*

JOSHUA D. ANGRIST AND ALAN B. KRUEGER

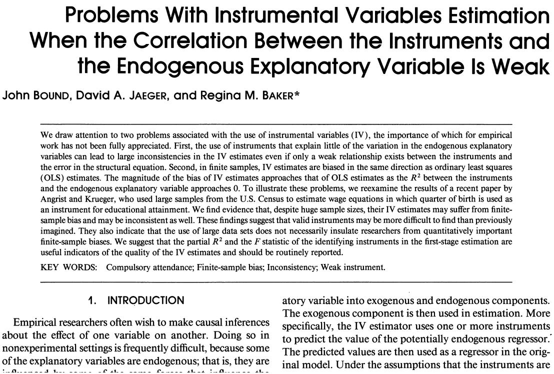
We establish that season of birth is related to educational attainment because of school start age policy and compulsory school attendance laws. Individuals born in the beginning of the year start school at an older age, and can therefore drop out after completing less schooling than individuals born near the end of the year. Roughly 25 percent of potential dropouts remain in school because of compulsory schooling laws. We estimate the impact of compulsory schooling on earnings by using quarter of birth as an instrument for education. The instrumental variables estimate of the return to education is close to the ordinary least squares estimate, suggesting that there is little bias in conventional estimates.

Every developed country in the world has a compulsory schooling requirement, yet little is known about the effect these laws have on educational attainment and earnings.[1] This paper exploits an unusual natural experiment to estimate the impact of compulsory schooling laws in the United States. The experiment

This example was mentioned earlier.

While n is huge, it turns out that the instrument is arguably weak.

Instrumental Variables



Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak

John BOUND, David A. JAEGER, and Regina M. BAKER*

We draw attention to two problems associated with the use of instrumental variables (IV), the importance of which for empirical work has not been fully appreciated. First, the use of instruments that explain little of the variation in the endogenous explanatory variables can lead to large inconsistencies in the IV estimates even if only a weak relationship exists between the instruments and the error in the structural equation. Second, in finite samples, IV estimates are biased in the same direction as ordinary least squares (OLS) estimates. The magnitude of the bias of IV estimates approaches that of OLS estimates as the R^2 between the instruments and the endogenous explanatory variable approaches 0. To illustrate these problems, we reexamine the results of a recent paper by Angrist and Krueger, who used large samples from the U.S. Census to estimate wage equations in which quarter of birth is used as an instrument for educational attainment. We find evidence that, despite huge sample sizes, their IV estimates may suffer from finite-sample bias and may be inconsistent as well. These findings suggest that valid instruments may be more difficult to find than previously imagined. They also indicate that the use of large data sets does not necessarily insulate researchers from quantitatively important finite-sample biases. We suggest that the partial R^2 and the F statistic of the identifying instruments in the first-stage estimation are useful indicators of the quality of the IV estimates and should be routinely reported.

KEY WORDS: Compulsory attendance; Finite-sample bias; Inconsistency; Weak instrument.

1. INTRODUCTION

Empirical researchers often wish to make causal inferences about the effect of one variable on another. Doing so in nonexperimental settings is frequently difficult, because some of the explanatory variables are endogenous; that is, they are influenced by some of the same forces that influence the

atory variable into exogenous and endogenous components. The exogenous component is then used in estimation. More specifically, the IV estimator uses one or more instruments to predict the value of the potentially endogenous regressor. The predicted values are then used as a regressor in the original model. Under the assumptions that the instruments are

These authors replicate some tables in the earlier paper with a random "instrument" they added to the data.

This spawned a large literature.
(The "Pitman drift" device from earlier slides follows Staiger and Stock (1997).)