Instrumental-Variable

Ruben Zamar
Department of Statistics UBC

March 12, 2014

THE LINEAR REGRESSION MODEL

- Ordinary linear regression model

$$\mathbf{y} = \mathbf{1}\alpha + X\beta + \varepsilon$$

THE LINEAR REGRESSION MODEL

- Ordinary linear regression model

$$\mathbf{y} = \mathbf{1}\alpha + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Design Matrix

$$X_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ x_{31} - \bar{x}_1 & x_{32} - \bar{x}_2 & \cdots & x_{3p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

THE LS ESTIMATOR

Ordinary Least Squares Estimate

$$\widehat{\beta}_{LS} = \left(X_c' X_c\right)^{-1} X_c' \mathbf{y} \quad , \quad \widehat{\alpha}_{LS} = \overline{y} - \overline{\mathbf{x}}' \widehat{\beta}_{LS}$$

Ordinary Least Squares Estimate

$$\widehat{\beta}_{LS} = \left(X_c' X_c\right)^{-1} X_c' \mathbf{y} \quad , \quad \widehat{\alpha}_{LS} = \overline{y} - \overline{\mathbf{x}}' \widehat{\beta}_{LS}$$

$$\widehat{\Sigma}_{xx} = \frac{1}{n} X_c' X_c \quad \text{(sample covariance matrix)}$$

$$\widehat{\Sigma}_{xy} = \frac{1}{n} X_c' \mathbf{y}$$

Ordinary Least Squares Estimate

$$\widehat{\beta}_{LS} = \left(X_c' X_c\right)^{-1} X_c' \mathbf{y} \quad , \quad \widehat{\alpha}_{LS} = \overline{y} - \overline{\mathbf{x}}' \widehat{\beta}_{LS}$$

$$\begin{aligned} \widehat{\Sigma}_{xx} &= \frac{1}{n} X_c' X_c \quad \text{(sample covariance matrix)} \\ \widehat{\Sigma}_{xy} &= \frac{1}{n} X_c' \mathbf{y} \end{aligned}$$

$$\widehat{\beta}_{LS} = \widehat{\Sigma}_{xx}^{-1} \widehat{\Sigma}_{xy}$$

KEY ASSUMPTIONS

- All covariates are **exogenous**

$$X \quad \text{and} \quad \varepsilon \quad \text{are independent}$$

KEY ASSUMPTIONS

- All covariates are **exogenous**

$$X \quad \text{and} \quad \varepsilon \quad \text{are independent}$$

- Errors have **zero mean**

$$E(\varepsilon) \quad = \quad \mathbf{0}$$

LS IS UNBIASED

$$E\left\{\widehat{\beta}_{LS}\right\} = E\left\{E\left\{\widehat{\beta}_{LS}|X\right\}\right\}$$

LS IS UNBIASED

$$\begin{aligned} E\left\{\widehat{\beta}_{LS}\right\} &= E\left\{E\left\{\widehat{\beta}_{LS}|X\right\}\right\} \\ &= E\left\{\left(X_c'X_c\right)^{-1}X_c' \;\; E\left\{\mathbf{y}|X\right\}\right\} \end{aligned}$$

LS IS UNBIASED

$$
\begin{aligned}
E\left\{\widehat{\beta}_{LS}\right\} &= E\left\{E\left\{\widehat{\beta}_{LS}|X\right\}\right\} \\
&= E\left\{\left(X_c'X_c\right)^{-1}X_c' \;\; E\left\{\mathbf{y}|X\right\}\right\} \\
&= E\left\{\left(X_c'X_c\right)^{-1}X_c' \;\; E\left\{\mathbf{1}\alpha + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}|X\right\}\right\}
\end{aligned}
$$

LS IS UNBIASED

$$
\begin{aligned}
E\left\{\widehat{\beta}_{LS}\right\} &= E\left\{E\left\{\widehat{\beta}_{LS}|X\right\}\right\} \\
&= E\left\{\left(X_c'X_c\right)^{-1}X_c' \;\; E\left\{\mathbf{y}|X\right\}\right\} \\
&= E\left\{\left(X_c'X_c\right)^{-1}X_c' \;\; E\left\{\mathbf{1}\alpha + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}|X\right\}\right\}
\end{aligned}
$$

$$
= E\left\{\overbrace{\left(X_c'X_c\right)^{-1}X_c'\mathbf{1}}^{=0}\alpha + \overbrace{\left(X_c'X_c\right)^{-1}X_c'X}^{=I}\boldsymbol{\beta} + \left(X_c'X_c\right)^{-1}X_c'\overbrace{E\left\{\boldsymbol{\varepsilon}\right\}}^{=0}\right\}
$$

LS IS UNBIASED

$$
\begin{aligned}
E\left\{\widehat{\beta}_{LS}\right\} &= E\left\{E\left\{\widehat{\beta}_{LS}|X\right\}\right\} \\
&= E\left\{\left(X_c'X_c\right)^{-1}X_c' \;\; E\left\{\mathbf{y}|X\right\}\right\} \\
&= E\left\{\left(X_c'X_c\right)^{-1}X_c' \;\; E\left\{\mathbf{1}\alpha + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}|X\right\}\right\}
\end{aligned}
$$

$$
= E\left\{\overbrace{\left(X_c'X_c\right)^{-1}X_c'\mathbf{1}}^{=0}\alpha + \overbrace{\left(X_c'X_c\right)^{-1}X_c'X}^{=I}\boldsymbol{\beta} + \left(X_c'X_c\right)^{-1}X_c'\overbrace{E\left\{\boldsymbol{\varepsilon}\right\}}^{=0}\right\}
$$

$$
= \boldsymbol{\beta}
$$

LS IS CONSISTENT

$$\widehat{\beta}_{LS} = \widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xy} \longrightarrow \Sigma_{xx}^{-1}\Sigma_{xy}, \quad \text{by LLN}$$

LS IS CONSISTENT

$$\widehat{\beta}_{LS} = \widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xy} \longrightarrow \Sigma_{xx}^{-1}\Sigma_{xy}, \quad \text{by LLN}$$

$$\Sigma_{xx}^{-1}\Sigma_{xy} = \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, y\right)$$

LS IS CONSISTENT

$$\widehat{\beta}_{LS} = \widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xy} \longrightarrow \Sigma_{xx}^{-1}\Sigma_{xy}, \quad \text{by LLN}$$

$$\begin{aligned} \Sigma_{xx}^{-1}\Sigma_{xy} &= \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, y\right) \\ &= \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, \alpha + \mathbf{x}'\beta + \varepsilon\right) \end{aligned}$$

LS IS CONSISTENT

$$\widehat{\beta}_{LS} = \widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xy} \longrightarrow \Sigma_{xx}^{-1}\Sigma_{xy}, \quad \text{by LLN}$$

$$\begin{aligned}
\Sigma_{xx}^{-1}\Sigma_{xy} &= \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, y\right) \\
&= \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, \alpha + \mathbf{x}'\beta + \varepsilon\right) \\
&= \Sigma_{xx}^{-1}\overbrace{Cov\left(\mathbf{x}, \mathbf{x}'\beta\right)}^{\Sigma_{xx}\beta} + \Sigma_{xx}^{-1}\overbrace{Cov\left(\mathbf{x}, \varepsilon\right)}^{\mathbf{0}}
\end{aligned}$$

LS IS CONSISTENT

$$\widehat{\beta}_{LS} = \widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xy} \longrightarrow \Sigma_{xx}^{-1}\Sigma_{xy}, \quad \text{by LLN}$$

$$\begin{aligned}
\Sigma_{xx}^{-1}\Sigma_{xy} &= \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, y\right) \\
&= \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, \alpha + \mathbf{x}'\beta + \varepsilon\right) \\
&= \Sigma_{xx}^{-1} \overbrace{Cov\left(\mathbf{x}, \mathbf{x}'\beta\right)}^{\Sigma_{xx}\beta} + \Sigma_{xx}^{-1} \overbrace{Cov\left(\mathbf{x}, \varepsilon\right)}^{\mathbf{0}} \\
&= \beta
\end{aligned}$$

ENDOGENEITY

- Some covariates (called **endogenous**) are correlated with the error term

ENDOGENEITY

- Some covariates (called **endogenous**) are correlated with the error term
- This problem arises in several contexts:

- Some covariates (called **endogenous**) are correlated with the error term
- This problem arises in several contexts:

Errors-in-variables: Some covariates are measured with errors

ENDOGENEITY

- Some covariates (called **endogenous**) are correlated with the error term
- This problem arises in several contexts:

Errors-in-variables: Some covariates are measured with errors

Omitted covariates: Some model covariates are correlated with unoserved predictors

ENDOGENEITY

- Some covariates (called **endogenous**) are correlated with the error term
- This problem arises in several contexts:

Errors-in-variables: Some covariates are measured with errors

Omitted covariates: Some model covariates are correlated with unoserved predictors

Simultaneity: Some covariates simultaneously affect and are affected by the response variable

ENDOGENEITY

- The (additive) asymptotic bias due to endogeneity is given by

$$AB = \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, \varepsilon\right)$$

ENDOGENEITY

- The (additive) asymptotic bias due to endogeneity is given by

$$\begin{aligned} AB &= \Sigma_{xx}^{-1} Cov\left(\mathbf{x},\varepsilon\right) \\ &= \frac{Cov\left(x,\varepsilon\right)}{Var\left(x\right)} \quad \text{(single linear regression)} \end{aligned}$$

ENDOGENEITY

- The (additive) asymptotic bias due to endogeneity is given by

$$
\begin{aligned}
AB &= \Sigma_{xx}^{-1} Cov\left(\mathbf{x}, \varepsilon\right) \\
&= \frac{Cov\left(x, \varepsilon\right)}{Var\left(x\right)} \quad \text{(single linear regression)} \\
&= Corr\left(x, \varepsilon\right) \frac{sd\left(\varepsilon\right)}{sd\left(x\right)} \quad \text{(single linear regression)}
\end{aligned}
$$

A simple example:

$$\begin{aligned} y_i &= 1 + 2v_i + \varepsilon_i, \qquad i = 1, 2, ..., n \\ x_i &= v_i + e_i \ \ (\textbf{error-in-variable}) \end{aligned}$$

A simple example:

$$
\begin{aligned}
y_i &= 1 + 2v_i + \varepsilon_i, \qquad i = 1, 2, ..., n \\
x_i &= v_i + e_i \;\; (\textbf{error-in-variable})
\end{aligned}
$$

To fix ideas suppose that

$$v_i, \varepsilon_i \text{ and } e_i \text{ are independent}$$

$$v_i \sim N(0, 1), \qquad \varepsilon_i \, , \, e_i \sim N(0, 0.5)$$

EIV (continued)

$$\begin{aligned}
y_i &= 1 + 2v_i + \varepsilon_i \\
&= 1 + 2\left(x_i - e_i\right) + \varepsilon_i \\
&= 1 + 2x_i + \left(\varepsilon_i - 2e_i\right) \\
&= 1 + 2x_i + \Delta_i
\end{aligned}$$

$$\Delta_i = \varepsilon_i - 2e_i$$

EIV (continued)

$$\begin{aligned} y_i &= 1 + 2v_i + \varepsilon_i \\ &= 1 + 2\left(x_i - e_i\right) + \varepsilon_i \\ &= 1 + 2x_i + \left(\varepsilon_i - 2e_i\right) \\ &= 1 + 2x_i + \Delta_i \end{aligned}$$

$$\begin{aligned} \Delta_i &= \varepsilon_i - 2e_i \\ cov\left(x_i, \Delta_i\right) &= -2 \times 0.5 = -1 \quad (\textbf{endogeneity}) \\ Var\left(x_i\right) &= 1.5 \end{aligned}$$

EIV (continued)

$$
\begin{aligned}
y_i &= 1 + 2v_i + \varepsilon_i \\
&= 1 + 2\left(x_i - e_i\right) + \varepsilon_i \\
&= 1 + 2x_i + \left(\varepsilon_i - 2e_i\right) \\
&= 1 + 2x_i + \Delta_i
\end{aligned}
$$

$$
\begin{aligned}
\Delta_i &= \varepsilon_i - 2e_i \\
cov\left(x_i, \Delta_i\right) &= -2 \times 0.5 = -1 \quad (\textbf{endogeneity}) \\
Var\left(x_i\right) &= 1.5
\end{aligned}
$$

$$
\mathrm{AB} = \frac{cov\left(x_i, \Delta_i\right)}{Var\left(x_i\right)} = -\frac{1}{1.5} = -0.667
$$

Suppose that the "true" model is

$$y_i = \alpha + \beta_1 \overbrace{x_{1i}}^{\text{included}} + \beta_2 \overbrace{x_{2i}}^{\text{omitted}} + \overbrace{\varepsilon_i}^{\text{independent errors}}$$

OMITTED VARIABLES

Suppose that the "true" model is

$$y_i \quad = \quad \alpha + \beta_1 \overbrace{x_{1i}}^{\text{included}} + \beta_2 \overbrace{x_{2i}}^{\text{omitted}} + \overbrace{\varepsilon_i}^{\text{independent errors}}$$

To fix ideas suppose that

$$Var(x_{1i}) \quad = \quad Var(x_{2i}) = 1, \qquad var(\varepsilon_i) = 0.1$$

and

$$Cov(x_{1i}, x_{2i}) = 0.4,$$

OMITTED VARIABLES (continued)

Suppose variable x_2 is omitted and we fit the model

$$y_i = \alpha + \beta x_{1i} + \Delta_i$$

Suppose variable x_2 is omitted and we fit the model

$$
\begin{aligned}
y_i &= \alpha + \beta x_{1i} + \Delta_i \\
\Delta_i &= \beta_2 x_{2i} + \varepsilon_i \\
cov\left(\Delta_i, x_{1i}\right) &= 0.4\beta_2 \qquad (\textbf{endogeneity})
\end{aligned}
$$

OMITTED VARIABLES (continued)

Suppose variable x_2 is omitted and we fit the model

$$
\begin{aligned}
y_i &= \alpha + \beta x_{1i} + \Delta_i \\
\Delta_i &= \beta_2 x_{2i} + \varepsilon_i \\
cov\left(\Delta_i, x_{1i}\right) &= 0.4\beta_2 \qquad (\textbf{endogeneity})
\end{aligned}
$$

$$
\text{AB} = \frac{cov\left(x_{1i}, \Delta_i\right)}{Var\left(x_{1i}\right)} = 0.4\beta_2
$$

SOME EXAMPLES OF ENDOGENEITY

- **Risk factors of coronary heart disease**
  Carrol, Enciclopedy of Biostatistics, 2005: patients's long-term blood pressure may be measured with error (**errors-in-variables**)

SOME EXAMPLES OF ENDOGENEITY

- **Risk factors of coronary heart disease**
  Carrol, Enciclopedy of Biostatistics, 2005: patients's long-term blood pressure may be measured with error (**errors-in-variables**)

- **Effect of smoking on birth weight**
  Permutt and Hebel, Biometrics, 1989: smoking by pregnant mothers may be correlated with unobserved socioeconomic factors which affect their infant birth weight (**omitted covariates**)

SOME EXAMPLES OF ENDOGENEITY

- **Risk factors of coronary heart disease**
  Carrol, Enciclopedy of Biostatistics, 2005: patients's long-term blood pressure may be measured with error (**errors-in-variables**)

- **Effect of smoking on birth weight**
  Permutt and Hebel, Biometrics, 1989: smoking by pregnant mothers may be correlated with unobserved socioeconomic factors which affect their infant birth weight (**omitted covariates**)

- **Atherosclerotic cardiovascular disease**
  Holmes et al., PloS ONE, 2010: C-reactive protein (CPR) may **cause** atherogenesis or may **be produced** by inflammation within atherosclerotic plaque. Thus CPR may be a cause and a result of atherosclerosis (**simultaneity**)

THE METHOD OF INSTRUMENTAL VARIABLES

OVERVIEW

- First introduced in **Appendix B** of a book authored by **Philip G. Wright**, "TheTariff on Animal and Vegetable Oils", published in **1928**.

OVERVIEW

- First introduced in **Appendix B** of a book authored by **Philip G. Wright**, “TheTariff on Animal and Vegetable Oils”, published in **1928**.
  - Some people attribute the authorship of Appendix B to Philip’s eldest son, **Sewal Wright**

OVERVIEW

- First introduced in **Appendix B** of a book authored by **Philip G. Wright**, "TheTariff on Animal and Vegetable Oils", published in **1928**.
  - Some people attribute the authorship of Appendix B to Philip's eldest son, **Sewal Wright**
- The method consists of using "**instruments**" (variables not included in the model) to consistently estimate the **regression coefficients of endogenous covariates**

OVERVIEW

- First introduced in **Appendix B** of a book authored by **Philip G. Wright**, "TheTariff on Animal and Vegetable Oils", published in **1928**.
  - Some people attribute the authorship of Appendix B to Philip's eldest son, **Sewal Wright**
- The method consists of using "**instruments**" (variables not included in the model) to consistently estimate the **regression coefficients of endogenous covariates**
- Very useful to investigate possible **causal relations** in observational studies

INSTRUMENTAL VARIABLES

- "**Instruments**" are variables

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{qi} \end{pmatrix} \quad i = 1, 2, \ldots n, \qquad q \geq p$$

INSTRUMENTAL VARIABLES

- "**Instruments**" are variables

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{qi} \end{pmatrix} \quad i = 1, 2, \ldots n, \qquad q \geq p$$

  - $\mathbf{z}_i$ is **correlated** with the endogenous covariates

- "**Instruments**" are variables

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{qi} \end{pmatrix} \quad i = 1, 2, \ldots n, \qquad q \geq p$$

  - $\mathbf{z}_i$ is **correlated** with the endogenous covariates
  - $\mathbf{z}_i$ is **uncorrelated** with the error term

- "**Instruments**" are variables

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{qi} \end{pmatrix} \quad i = 1, 2, ...n, \qquad q \geq p$$

  - $\mathbf{z}_i$ is **correlated** with the endogenous covariates
  - $\mathbf{z}_i$ is **uncorrelated** with the error term
  - rank of $Cov(\mathbf{z}_i) = q$

INSTRUMENTAL VARIABLES

- "**Instruments**" are variables

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{qi} \end{pmatrix} \quad i = 1, 2, \ldots n, \qquad q \geq p$$

  - $\mathbf{z}_i$ is **correlated** with the endogenous covariates
  - $\mathbf{z}_i$ is **uncorrelated** with the error term
  - rank of $Cov(\mathbf{z}_i) = q$
  - rank of $Cov(\mathbf{z}_i, \mathbf{x}_i) = p$

THE CENTERED INSTRUMENT MATRIX

$$Z_c = \begin{pmatrix} z_{11} - z_1 & z_{12} - z_2 & \cdots & z_{1p} - \bar{z}_p \\ z_{21} - \bar{z}_1 & z_{22} - \bar{z}_2 & \cdots & z_{2p} - \bar{z}_p \\ z_{31} - \bar{z}_1 & z_{32} - \bar{z}_2 & \cdots & z_{3p} - \bar{z}_p \\ \vdots & \vdots & & \vdots \\ z_{n1} - \bar{z}_1 & z_{n2} - \bar{z}_2 & \cdots & z_{np} - \bar{z}_p \end{pmatrix}$$

THE METHOD

- First, project the model covariates on the space generated by the instruments

$$\widehat{X}_c = \overbrace{\left[Z_c \left(Z_c' Z_c \right)^{-1} Z_c' \right]}^{\text{projection matrix}} X_c = P_z X_c$$

THE METHOD

- First, project the model covariates on the space generated by the instruments

$$\widehat{X}_c = \overbrace{\left[Z_c \left(Z_c' Z_c \right)^{-1} Z_c' \right]}^{\text{projection matrix}} X_c = P_z X_c$$

- $\widehat{X}_c$ represents the part of X_c uncorrelated with ε

- Second, apply ordinary LS but using the design matrix $\widehat{X}_c$ instead of X_c :

- Second, apply ordinary LS but using the design matrix $\widehat{X}_c$ instead of X_c :

$$\begin{aligned}
\widehat{\beta}_{IV} &= \left(\widehat{X}_c'\widehat{X}_c\right)^{-1}\widehat{X}_c\mathbf{y} \\
&= \left(X_c'P_zX_c\right)^{-1}X_c'P_z\mathbf{y} \\
&= \left(X_c'Z_c\left(Z_c'Z_c\right)^{-1}Z_c'X_c\right)^{-1}X_c'Z_c\left(Z_c'Z_c\right)^{-1}Z_c'\mathbf{y}
\end{aligned}$$

- Second, apply ordinary LS but using the design matrix $\widehat{X}_c$ instead of X_c :

$$
\begin{aligned}
\widehat{\beta}_{IV} &= \left(\widehat{X}_c'\widehat{X}_c\right)^{-1}\widehat{X}_c\mathbf{y} \\
&= \left(X_c'P_zX_c\right)^{-1}X_c'P_z\mathbf{y} \\
&= \left(X_c'Z_c\left(Z_c'Z_c\right)^{-1}Z_c'X_c\right)^{-1}X_c'Z_c\left(Z_c'Z_c\right)^{-1}Z_c'\mathbf{y} \\
&= \left(\widehat{\Sigma}_{xz}\widehat{\Sigma}_{zz}^{-1}\widehat{\Sigma}_{zx}\right)^{-1}\widehat{\Sigma}_{xz}\widehat{\Sigma}_{zz}^{-1}\widehat{\Sigma}_{zy}
\end{aligned}
$$

- Second, apply ordinary LS but using the design matrix $\widehat{X}_c$ instead of X_c :

$$
\begin{aligned}
\widehat{\beta}_{IV} &= \left(\widehat{X}_c'\widehat{X}_c\right)^{-1}\widehat{X}_c\mathbf{y} \\
&= \left(X_c'P_zX_c\right)^{-1}X_c'P_z\mathbf{y} \\
&= \left(X_c'Z_c\left(Z_c'Z_c\right)^{-1}Z_c'X_c\right)^{-1}X_c'Z_c\left(Z_c'Z_c\right)^{-1}Z_c'\mathbf{y} \\
&= \left(\widehat{\Sigma}_{xz}\widehat{\Sigma}_{zz}^{-1}\widehat{\Sigma}_{zx}\right)^{-1}\widehat{\Sigma}_{xz}\widehat{\Sigma}_{zz}^{-1}\widehat{\Sigma}_{zy} \qquad \Longleftarrow \text{KEY OBSERVATION}
\end{aligned}
$$

STATISTICAL PROPERTIES

- $\widehat{\beta}_{IV}$ is **consistent** and **asymptotically normal** under mild regularity assumptions

STATISTICAL PROPERTIES

- $\widehat{\beta}_{IV}$ is **consistent** and **asymptotically normal** under mild regularity assumptions
- But data may contain outliers in the
  - **response variable** y_i
  - **covariates** $\mathbf{x}_i$
  - **instruments** $\mathbf{z}_i$

STATISTICAL PROPERTIES

- $\widehat{\beta}_{IV}$ is **not robust** (sensitive to outliers in the response, covariates and/or instruments)

STATISTICAL PROPERTIES

- $\widehat{\beta}_{IV}$ is **not robust** (sensitive to outliers in the response, covariates and/or instruments)
    - $\widehat{\beta}_{IV}$ has **unbounded influence** function

STATISTICAL PROPERTIES

- $\widehat{\beta}_{IV}$ is **not robust** (sensitive to outliers in the response, covariates and/or instruments)
  - $\widehat{\beta}_{IV}$ has **unbounded influence** function
  - $\widehat{\beta}_{IV}$ has **zero breakdown point**

STATISTICAL PROPERTIES

- $\widehat{\beta}_{IV}$ is **not robust** (sensitive to outliers in the response, covariates and/or instruments)
  - $\widehat{\beta}_{IV}$ has **unbounded influence** function
  - $\widehat{\beta}_{IV}$ has **zero breakdown point**
  - $\widehat{\beta}_{IV}$ **fails robustness-tests** in simulations

STATISTICAL PROPERTIES

- $\widehat{\beta}_{IV}$ is **not robust** (sensitive to outliers in the response, covariates and/or instruments)
  - $\widehat{\beta}_{IV}$ has **unbounded influence** function
  - $\widehat{\beta}_{IV}$ has **zero breakdown point**
  - $\widehat{\beta}_{IV}$ **fails robustness-tests** in simulations
  - $\widehat{\beta}_{IV}$ **has poor performance in some real data examples** due to outliers.

APPLICATION OF INSTRUMENTAL VARIABLES IN MEDICAL RESEARCH

- Prior work suggests that **high blood pressure** can increase the size of the **left atrial dimension**
(Vaziri et al., 1995; Milan et al., 2009; Benjamin et al., 1995).

- Prior work suggests that **high blood pressure** can increase the size of the **left atrial dimension** (Vaziri et al., 1995; Milan et al., 2009; Benjamin et al., 1995).
- We use IV to measure the effect of *long-term systolic blood pressure* (SBP) on left atrial size (LAD).

DATA

- Our sample includes **164** male subjects from the original cohort who underwent their 16th biennial examination and satisfy the inclusion criteria employed by Vaziri et al. (1995):
  - patient doesn't have a history of heart disease,
  - patient is not taking cardiovascular medication
  - patient has complete clinical data

STATISTICAL MODEL

Following Vaziri et al. (1995), we consider the following model:

Variable	Description	Variable Status
LAD	Left atrial dimension	**Response**
SBP$_{16}$	long-term systolic blood pressure (Examination # 16)	Main covariate **Endogenous (measured with error)**
BMI	Body mass index	Control covariate **Exogenous**
AGE	Subject's age	Control covariate **Exogenous**
SBP$_{15}$	long-term systolic blood pressure (Examination # 15)	**Instrument**

RESULTS

	SBP	BMI	AGE
COEFF.	0.011	**3.195**	0.023
P-VALUE	0.684	<0.001	0.685

APPLICATION OF INSTRUMENTAL VARIABLES IN LABOR ECONOMICS

APPLICATION OF RIV TO LABOR UNIONS DATA

- Our dataset contains 181 industries in the United States for the year 2003.

APPLICATION OF RIV TO LABOR UNIONS DATA

- Our dataset contains 181 industries in the United States for the year 2003.
- Unionization rates are obtained from the Union Membership and Coverage Database

APPLICATION OF RIV TO LABOR UNIONS DATA

- Our dataset contains 181 industries in the United States for the year 2003.
- Unionization rates are obtained from the Union Membership and Coverage Database
- The fraction of male workers comes from the Census Population Survey

APPLICATION OF RIV TO LABOR UNIONS DATA

- Our dataset contains 181 industries in the United States for the year 2003.
- Unionization rates are obtained from the Union Membership and Coverage Database
- The fraction of male workers comes from the Census Population Survey
- The remaining variables are constructed using data from Compustat

APPLICATION OF RIV TO LABOR UNIONS DATA

- Our dataset contains 181 industries in the United States for the year 2003.
- Unionization rates are obtained from the Union Membership and Coverage Database
- The fraction of male workers comes from the Census Population Survey
- The remaining variables are constructed using data from Compustat

EFFECT OF LABOR UNIONS ON FIRMS' PROFITS

Variable Name	Variable Description	Variable Status
ROA	Accounting profits scaled by assets	**Response variable**
UNION	Fraction of workers that are unionized	**Endogenous**, main covariate
GROWTH	Sales growth rate	**Exogenous,** control covariate
CAPEX	Capital expenditures scaled by assets	**Exogenous**, control covariate

EFFECT OF LABOR UNIONS ON FIRMS' PROFITS

Variable Name	Variable Description	Variable Status
SIZE	Fixed assets (log scale)	**Exogenous**, control covariate
LEV	Total debt scaled by assets	**Exogenous**, control covariate
KL	Capital-labor ratio	**Exogenous**, control covariate
MALE	Fraction of male workers	**Instrument**

UNION

- Why is UNION endogenous?

UNION

- Why is UNION endogenous?
- There are two possible reasons

UNION

- Why is UNION endogenous?
- There are two possible reasons
  1. UNION and ROA are likely to be simultaneously determined

UNION

- Why is UNION endogenous?
- There are two possible reasons
  1. UNION and ROA are likely to be simultaneously determined
     - Industry higher performance may stimulate unionization to get a better profits share

UNION

- Why is UNION endogenous?
- There are two possible reasons
  1. UNION and ROA are likely to be simultaneously determined
     - Industry higher performance may stimulate unionization to get a better profits share
     - It is conjectured that higher unionization cause industry performance to decline

UNION

- Why is UNION endogenous?
- There are two possible reasons
  1. UNION and ROA are likely to be simultaneously determined
     - Industry higher performance may stimulate unionization to get a better profits share
     - It is conjectured that higher unionization cause industry performance to decline
  2. There may be unobserved industry features that determine performance and are correlated with unionization

MALE

- Why is MALE a reasonable instrument for UNION?

MALE

- Why is MALE a reasonable instrument for UNION?
  - males have more permanent attachment to the labor market and to internal job ladders

MALE

- Why is MALE a reasonable instrument for UNION?
  - males have more permanent attachment to the labor market and to internal job ladders
    $\Rightarrow$ males derive higher wage/non-wage benefits from unionizing

MALE

- Why is MALE a reasonable instrument for UNION?
  - males have more permanent attachment to the labor market and to internal job ladders
    $\Rightarrow$ males derive higher wage/non-wage benefits from unionizing
    $\Rightarrow$ males are more likely to become unionized.

RESULTS

	UNION	GROWTH	CAPEX	SIZE	LEV	KL
coeff.	-0.271	**0.126**	**0.197**	0.003	**0.094**	0.004
p-value	0.128	0.013	0.000	0.312	0.005	0.313