



Centre for Market and
Public Organisation

Instrumental variable estimation for binary outcomes

Paul Clarke & Frank Windmeijer

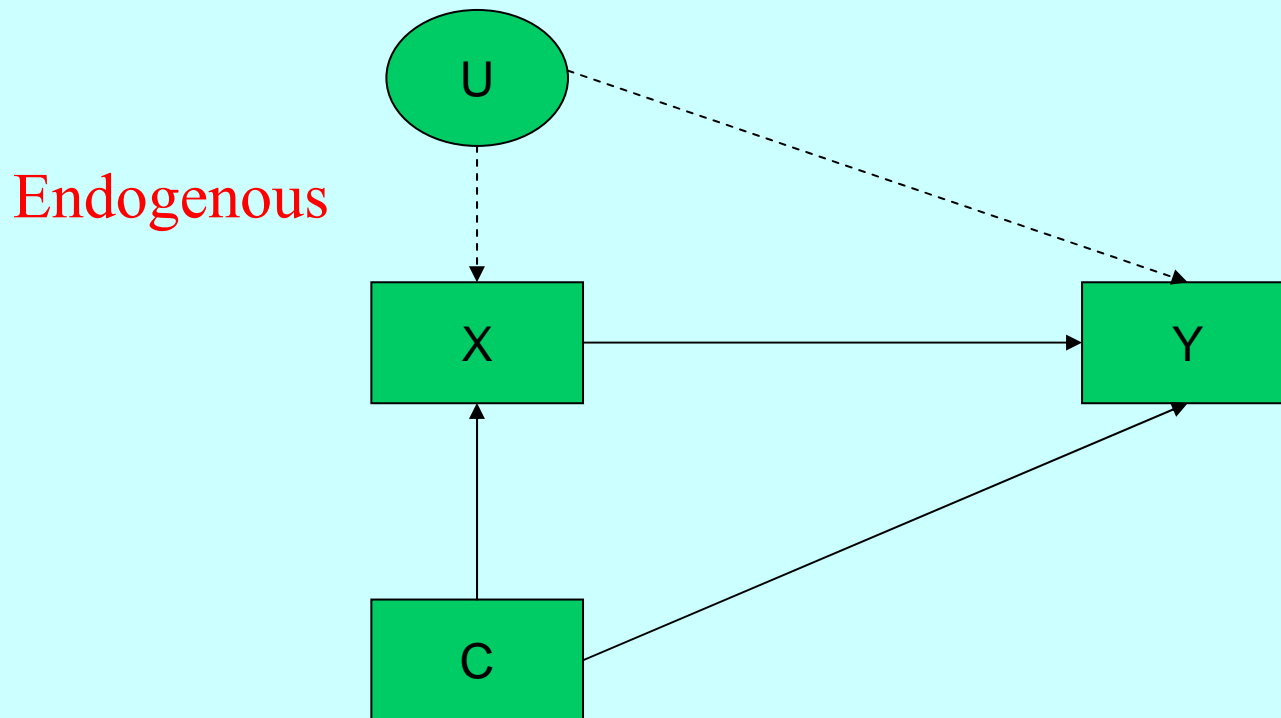
Presented at joint meeting of RSS General Applications and Social Statistics
sections “Causality in Statistical Investigations” 27 May 2009



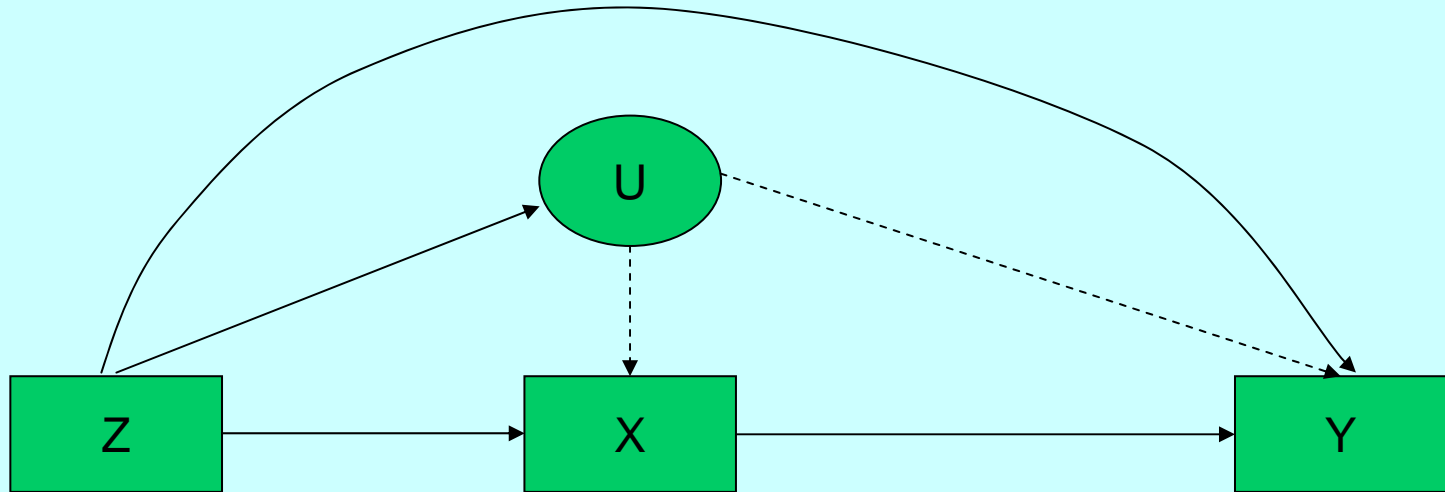
University of
BRISTOL

The Leverhulme Trust

A Problem Of Omitted Variables

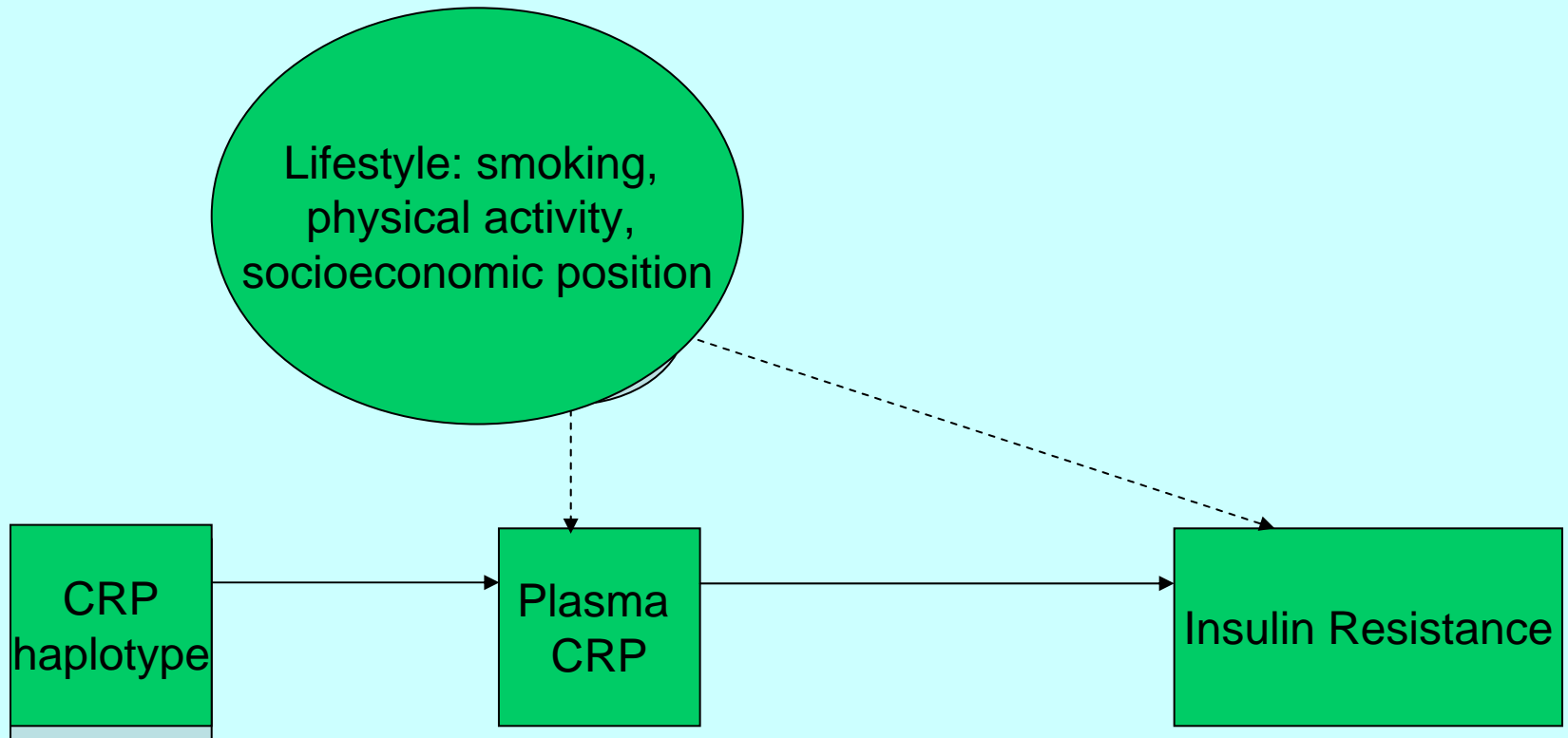


An Instrumental Variable (IV)



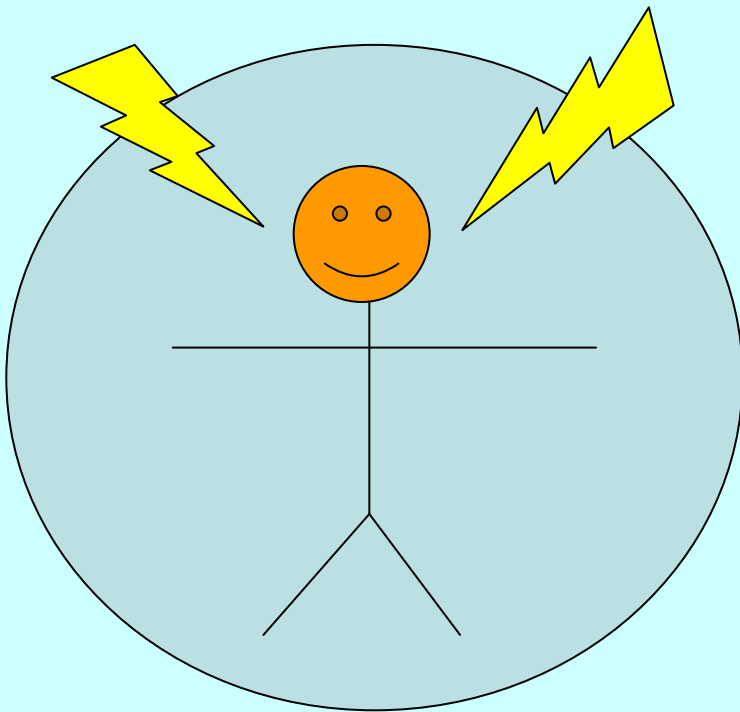
- 3 Core Conditions:
1. Z predicts X
 2. $Z \perp\!\!\!\perp Y \mid X, U$
 3. $Z \perp\!\!\!\perp U$

Mendelian Randomisation

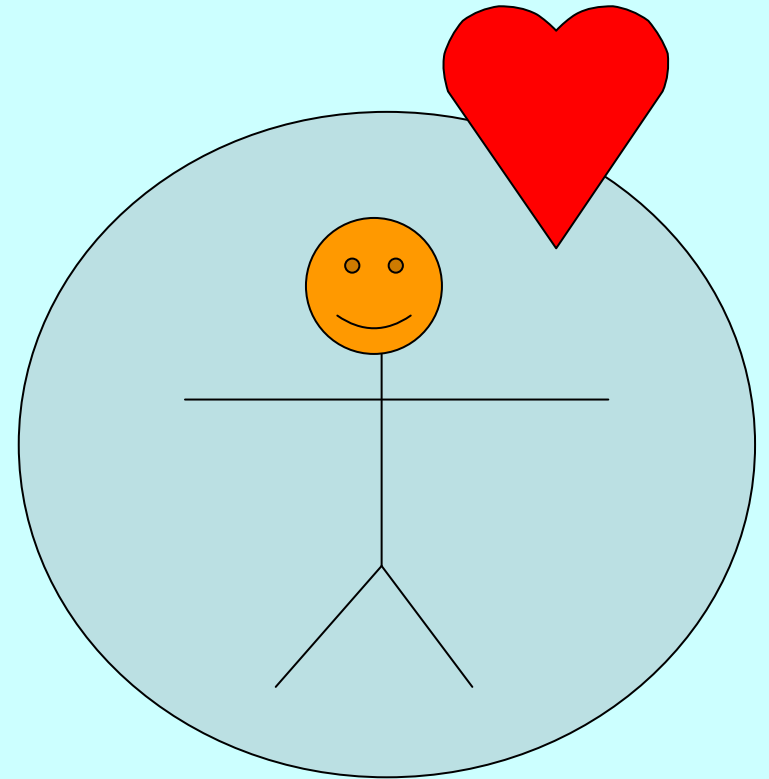


Potential Outcomes

$Y(1)$ ← For binary exposure → $Y(0)$



Exposed (high stress)



Unexposed (low stress)

Structural Models

- Model for potential outcomes
 - i.e. what would have happened
- Example: linear structural model

$$Y(x) = \beta_0 + x\beta_1 + U$$

- Average causal/treatment effect

$$\beta_1 = E(Y(1) - Y(0))$$

Descriptive Model

- Fit regression model to observed data

$$E(Y_i | X_i = x_i) = b_0 + x_i b_1$$

- OLS estimator is inconsistent

$$\hat{b}_1 \rightarrow \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)}$$

Linear IV Estimator

$$\hat{\beta}_{\text{IV}} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} \rightarrow \beta_1$$

Generalisation

- Two-stage least squares (2SLS)
 - Stage 1: Fit linear ‘reduced-form’ model
 - Stage 2: Replace X with its fitted value

$$x_i = a + z_i b + v_i \Rightarrow \hat{X}_i = \hat{a} + z_i \hat{b}$$

Binary Y : Where is U ?

- Binary structural model

$$Y(x) = \begin{cases} 1 & \text{if } \beta_0 + x\beta_1 + U > 0 \\ 0 & \text{if } \beta_0 + x\beta_1 + U \leq 0 \end{cases} = I(\beta_0 + x\beta_1 + U > 0)$$

- If U is standard logistic/normal

$$\Pr(Y(x) = 1) = E_U \{I(U > -\beta_0 - x\beta_1)\} = \begin{cases} \frac{\exp(\beta_0 + x\beta_1)}{1 + \exp(\beta_0 + x\beta_1)} \\ \Phi(\beta_0 + x\beta_1) \end{cases}$$

Logistic Case

- Causal odds ratio (COR)

$$\exp(\beta_1) = \frac{\Pr(Y(1) = 1)/\Pr(Y(1) = 0)}{\Pr(Y(0) = 1)/\Pr(Y(0) = 0)}$$

- Descriptive model

$$\text{logit } \Pr(Y_i = 1 | X_i = x_i) \neq \beta_0 + x_i \beta_1$$

Identification

- Chesher¹ shows:
 - Core conditions & binary structural model do not (point) identify β_1
 - Need conditional distribution of U given X

¹Chesher (2008) Endogeneity and discrete outcomes. *CeMAPP Working Paper 30/08*, University College London.

Maximum Likelihood

- Reduced-form model for X given Z
 - **Key** to identification (c.f. 2SLS)
 - Different assumptions \Rightarrow Different estimates
- ML for probit models (Rivers & Vuong¹)
 - i.e. assume $U \sim N(0, 1)$
 - Fit in Stata (`ivprobit`) or Mplus
 - 2-stage estimators available

¹Rivers & Vuong (1988) Limited information estimators and exogeneity tests, *J. Econometrics* pp 347-66

Semi-parametric Models

- Potential outcomes approach
 - Semi-parametric structural model
 - i.e. no appearance of U
 - Widely used in biostatistics
- Randomised controlled trials
 - Non-compliance: patients defy randomisation
 - Non-ignorable: depends on outcome

Marginal Estimator

- Ten Have¹ use Marginal Structural Models

$$\text{logit Pr}(Y(x) = 1) = \beta_0 + x\beta_1$$

- Estimating equation consistent if¹

$$E\{(Z - E(Z))\tilde{U}\} = E(\tilde{U}) = 0$$

$$\tilde{u}_i = y_i - \frac{\exp(\beta_0 + x_i\beta_1)}{1 + \exp(\beta_0 + x_i\beta_1)}$$

¹Ten Have, Joffe, Cary (2003) Causal logistic models for non-compliance. *Statist. Med.* pp 1255-83

Mean Separability

- Linear model is mean separable:

$$U = Y(x) - \beta_0 - x\beta_1 = Y - E(Y(x))$$

- But binary model is not

$$Y(x) = I(\beta_0 + x\beta_1 + U > 0) \\ \neq \frac{\exp(\beta_0 + x\beta_1)}{1 + \exp(\beta_0 + x\beta_1)} + \tilde{U}$$

Structural Mean Model

- Estimate OR among exposed groups¹

$$\text{logit Pr}(Y = 1|X = 1, Z = z) - \text{logit Pr}(Y(0) = 1|X = 1, Z = z) = \psi_z$$

- For identification:
 - No effect modification: $\exp \psi_z = \exp \psi$
 - Equals COR if ignorable non-compliance
- Holds only for special case
 - Placebo-control design/treatment exclusion

¹Vansteelandt, Goetghebeur (2003) Causal inference with generalized SMMs, *J. R. Statist. Soc. B* pp 817-35

Simulation Study

$$U \sim \text{logistic}(0,1)$$

$$Z \sim \text{Bernoulli}(0.5)$$

$$X|Z = z, U = u \sim \text{Bernoulli}\{\text{expit}(0.05z + \omega u)\}$$

$$Y = I(\beta_0 + x + u > 0)$$

-
- 100 simulations of $n = 1000$
 - Weak instrument (its effect on X is small)
 - True causal log-odds ratio is $\exp(1) = 2.7$

Results

Event prob.	50% ($\beta_0 = 0$)		5% ($\beta_0 = -3$)	
	$\omega = 0.5$	$\omega = 2.0$	$\omega = 0.5$	$\omega = 2.0$
Naïve	1.03 (0.2)	3.54 (0.5)	0.90 (0.3)	1.29 (0.3)
Marginal	-0.15 (0.3)	-0.47 (0.2)	0.34 (0.4)	0.58 (0.4)
SMM	0.15 (0.3)	1.19 (0.8)	0.01 (0.4)	0.12 (0.3)
ML Probit	0.09 (0.3)	0.81 (0.5)	-1.37 (3)	-13.6 (3)

Presented: bias and estimated standard error (Monte Carlo estimate)

Conclusions

- More assumptions than for linear models
- ML estimators weakly identified
 - Bayesian methods too
 - Sensitive to assumptions
- Semi-parametric estimators approximate
- Robust estimation of ‘local’ effects:
 - Other SMMs (Hernán & Robins, 2006)
 - Local average response models (Abadie, 2003)

- For further details see:
 - Clarke & Windmeijer (2009a) “Instrumental variable estimators for binary outcomes” *CMPO Working paper 09/209*, University of Bristol
(<http://www.bris.ac.uk/cmpo/publications/papers/>)
 - Clarke & Windmeijer (2009b) “Identification of local causal effects on binary outcomes using structural mean models” (working paper online soon).