Inference without models, II
EITM 2007

Jeffrey B. Lewis

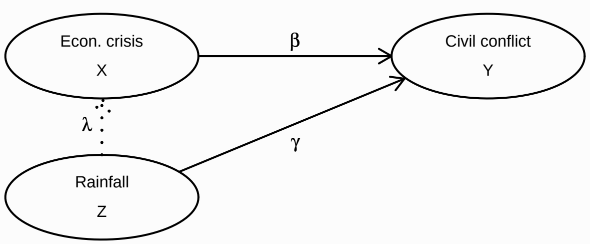
University of California, Los Angeles

July 3, 2007

Instrumental variables

- An *instrument* is a variable that is correlated with your independent variable, but has no direct affect on your dependent variable.
- One classic example is the use of draft number as an instrument for military service in models of income (Angrist 1990).
- The best instruments are obviously exogenous (birthdate, weather, random assignment).
- Unfortunately, often hard to come by.

Understanding why IV works

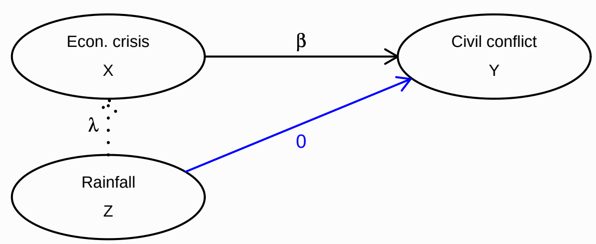


$$\mathrm{E}(X|Z) = \kappa + \lambda Z$$

$$\mathrm{E}(Y|Z) = \alpha + (\gamma + \lambda\beta)Z$$

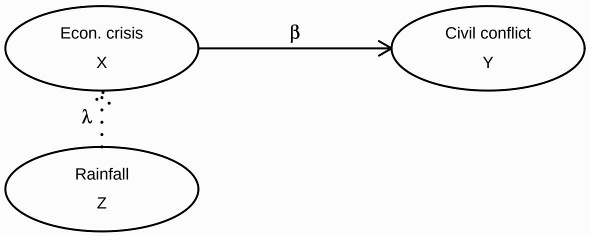
$$\mathrm{E}(Y|Z) = \alpha + (0 + \lambda\beta)Z$$

Understanding why IV works



$$\mathrm{E}(X|Z) = \kappa + \lambda Z$$

Understanding why IV works



$$\mathrm{E}(X|Z) = \kappa + \lambda Z$$

$$\mathrm{E}(Y|Z) = \alpha + (\lambda\beta)Z$$

The Simplest IV Estimator

1. Estimate
$$\mathrm{E}(Y|Z) = \alpha + (\lambda\beta)Z$$
by OLS and obtain an estimate of $\lambda\beta$, $\widehat{\lambda\beta}$.
2. Estimate
$$\mathrm{E}(X|Z) = \kappa + \lambda Z$$
by OLS and obtain an estimate of λ, $\widehat{\lambda}$.
3. Estimate β as
$$\frac{\widehat{\lambda\beta}}{\widehat{\lambda}}$$

IV implies the idea of a structural model

We assert that
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$
is the process that in some sense *generates* Y.

- Bias in the estimation of this regression is caused by correlations between the Xs and ϵ.
- IV works because:
  1. Z (the instrument) is uncorrelated with the error, ϵ, in the structural equation.
  2. Z does not *belong* in the structural equation.

Theory has made its way in!

Economic crisis and civil war

Miguel, Satyanath, and Sergenti (2004) consider effect of GDP growth on civil war.

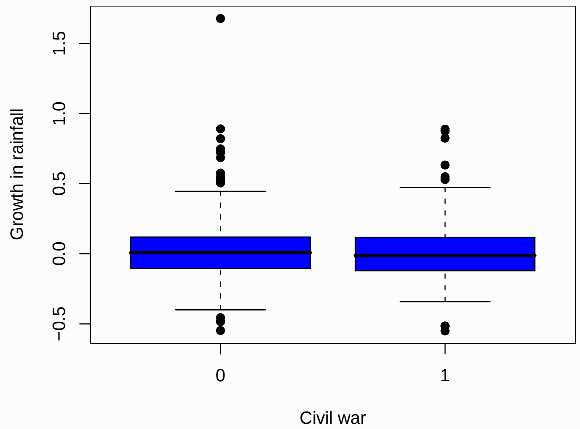
In particular, they consider the need for an instrumental variable.

- Endogeneity: conflict affects the economy.
- Omitted variables: Institutional factors
- Measurement error: GDP data unreliable

Instrument they choose is *rainfall*.

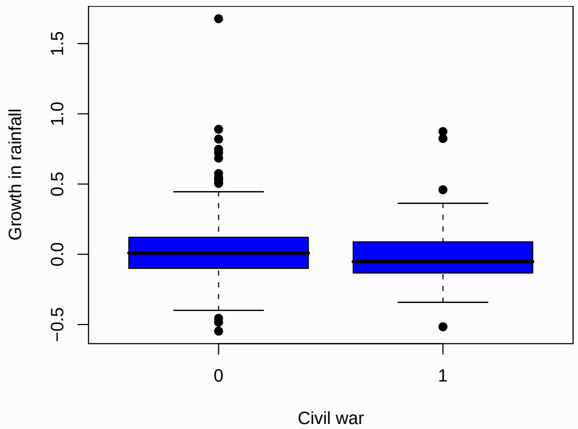
Limit sample to sub-saharan Africa in 1981-1999.

Economic crisis and civil war



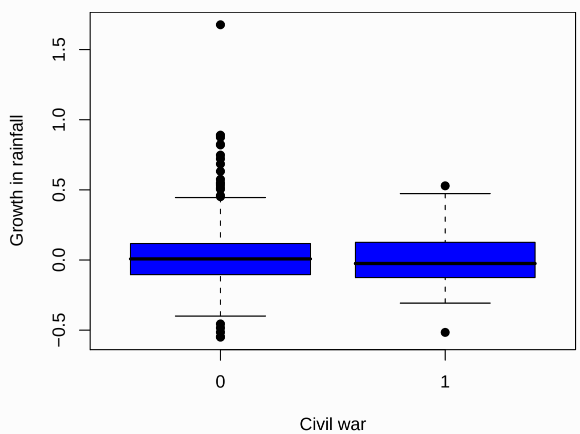
	Rainfall growth
$<$ 25 Battle death	0.017
$\geq$ 25 Battle deaths	0.021
Difference	0.004

Economic crisis and civil war



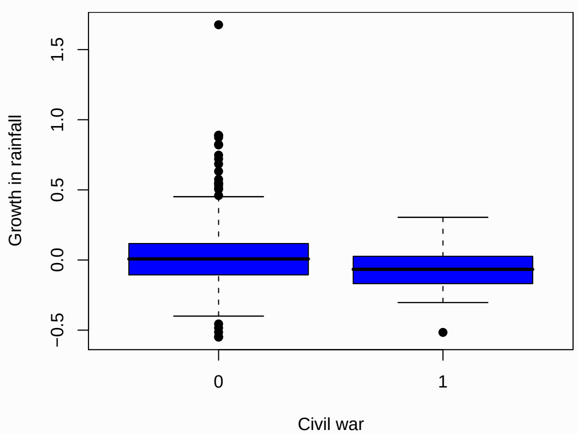
	Rainfall growth
$<$ 25 Battle death | No onset	0.019
$\geq$ 25 Battle deaths | Onset	0.022
Difference	0.003

Economic crisis and civil war



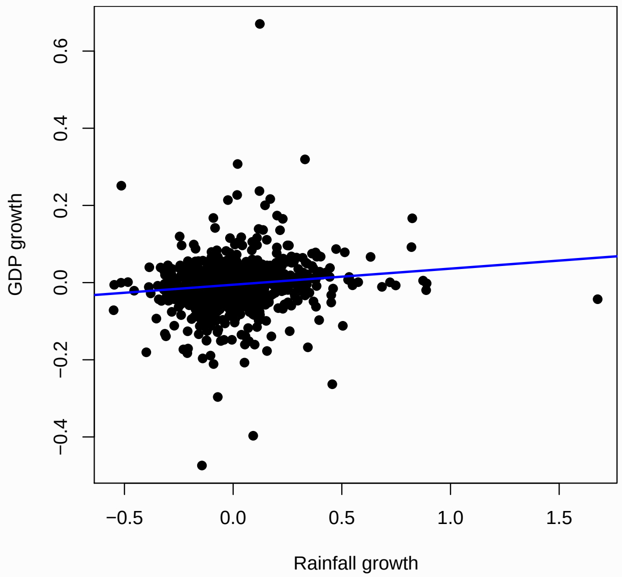
	Rainfall growth
$<$ 1000 Battle deaths	0.021
$\geq$ 1000 Battle deaths	0.002
Difference	-0.019

Economic crisis and civil war



	Rainfall growth
$<$ 1000 Battle death | No onset	0.020
$\geq$ 1000 Battle deaths | Onset	-0.072
Difference	-0.092

Economic crisis and civil war



Economic crisis and civil war

Reduced form equation:
Dependent variable = GDP Growth

	(1)	(2)
Growth in rainfall (t)	0.055	0.040
	(0.016)	(0.017)
Growth in rainfall ($t-1$)	0.054	0.028
	(0.013)	(0.014)
Country time trends	No	Yes
Country fixed effects	No	Yes
R^2	0.02	0.06

Economic crisis and civil war

Reduced form equation:
Dependent variable = Civil War

	($BD \geq 25$)	($BD \geq 1000$)
Growth in rainfall (t)	-0.024	-0.062
	(0.048)	(0.030)
Growth in rainfall ($t-1$)	-0.122	-0.069
	(0.052)	(0.052)
Country time trends	Yes	Yes
Country fixed effects	Yes	Yes
R^2	0.71	0.70

Economic crisis and civil war

Structural equation:
Dependent variable = Civil War

	$DB \geq 25$		$BD \geq 1000$
	(OLS)	(IV/2SLS)	(IV/2SLS)
Econ. Growth (t)	-0.21	-1.13	-1.48
	(1.48)	(1.40)	(0.82)
Econ. Growth ($t-1$)	0.07	-2.55	-0.77
	(0.16)	(1.16)	(0.76)
Country time trends	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes

Models of Selection

- Problems of selection plague inference in strategic settings.
- In some sense, selection is a really a problem of omitted variables. Actors condition their behavior on things that we, as analysts, do not directly observe.
- In order to address selection problems in the study of conflict, we need not only data on the outcomes of conflicts, but also data on potential conflicts that never occurred.
- Correcting for selection involves modelling the phenomena we are trying to explain and the process which generated the sample of data that we observe.

The Heckman selection model

- The outcome equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- The selection equation

$$S^* = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Z + \nu$$

where on observation is selected into the outcome equation sample if $S^* > 0$.

The Heckman selection model

Outcome: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
Selection: $S^* = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Z + \nu$

An important quantity is ρ, the correlation between ϵ and ν.

- If $\rho = 0$, the selection is only a function of independent variables (X_1 and X_2) and things that are independent of Y (Z and ν).
- If $\rho \neq 0$ (and either $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$), then selection induces a correlation between the Xs and ϵ in the outcome equation.

Estimating the Heckman selection model

The simple "two-step" approach:

- Assume ϵ and ν are bivariate normal.
- Estimate selection equation by probit.
- Estimate the outcome equation by OLS controlling for $\mathrm{E}(\epsilon | S^* > 0, X_1, X_2, Z)$.

Applying the standard selection model

$$S^* = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Z + \nu$$

Letting

$$\mu = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Z$$

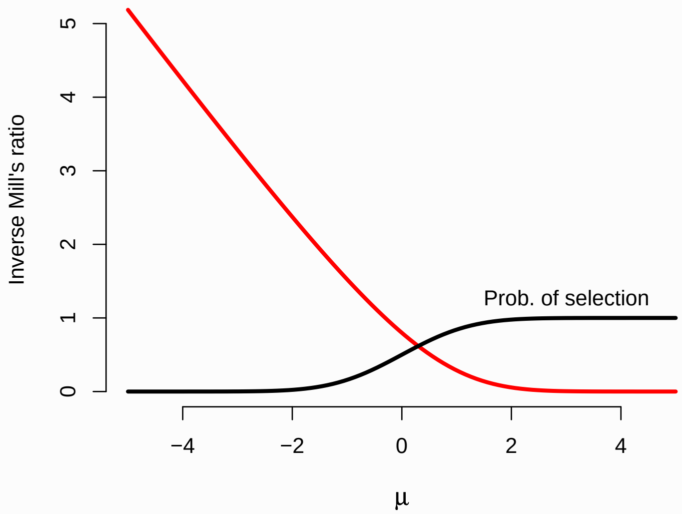
we have

$$\mathrm{E}(\epsilon | S^* > 0) = \rho\sigma \left(\frac{\phi(-\mu)}{1 - \Phi(-\mu)} \right)$$

where ϕ and Φ are the standard normal PDF and CDF.

$\frac{\phi(x)}{1-\Phi(x)}$ is known as the "inverse Mills ratio."

The inverse Mills ratio



Second stage regression

After fitting the first-stage probit, then fit

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_\lambda \left(\frac{\phi(-\hat{\mu})}{1 - \Phi(-\hat{\mu})} \right) + \delta$$

1. ρ can be inferred from $\hat{\lambda}$ and $\hat{\sigma}$.
2. Note that without Z, the model is only identified from non-linearity.
3. Problem of non-identification (or near non-identification) even more severe when the second stage is also probit (outcome is modelled also discrete).

Selection models in IR

- Real need for restrictions ("instruments") which may be theoretically hard to motivate.
- Results reliant on distributional assumptions and functional forms.
- Problems made more difficult by discrete outcome equations.
- Extreme value of ρ are often encountered.
- Must take great care.

IMF Programs and growth

Considered by Przeworkski and Vreeland (2000).
Selection:

$$S^* = \alpha_0 + \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \nu$$

Cases with an IMF agreement:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Cases with no IMF agreement:

$$Y' = \beta_0' + \beta_1' X_1 + \beta_2' X_2 + \epsilon'$$

In an IMF agreement if $S^* > 0$, otherwise not in an agreement and $(\nu, \epsilon, \epsilon')$ are multivariate normal.

IMF agreements and growth

Effect they would like to measure is $\mathrm{E}(Y - Y')$, but this cannot be directly observed.

Using there selection models' predictions What they can estimate is

$$\mathrm{E}(Y|X_1, X_2) - \mathrm{E}(Y'|X_1, X_2)$$

which accounts for difference in observable and non-observable growth determinants.

IMF agreements and growth

	Mean growth	
	Observed	Corrected
Under agreement	2.04	2.00
Not under agreement	4.39	3.66

IMF agreements and growth

Actually, P&V is a little more complicated...
Their selection stage is not a simple probit.

		IMF	
		Yes	No
Country	Yes	G_{yy}	G_{yn}
	No	G_{ny}	G_{nn}

They suppose that there are two latent selection variables, S^*_{IMF} and $S^*_{Country}$, both of which must be greater than 0 to observe an agreement. We will discuss this model in more detail tomorrow.

Recap

- Standard econometric solutions to problems of endogeneity, selection, and so forth can and are applied to the strategic setting in IR.
- The application of these models, involves the use of particular empirical models the specification of which can be informed by theory. Ultimately, some hypotheses must be maintained.
- The standard modelling assumptions will generally only approximate the relationships implied by formal theory.
- In the next two sessions, we will consider how theory can be more directly integrated into estimation.
- But, there are important caveats and trade-offs.