From Local to Global: A Case Study in External Validity

Rajeev Dehejia
New York University

joint work with

Crisitian Pop-Eleches
Columbia University

Cyrus Samii
New York University
Introduction

questions

Can an experiment from location A (B, C,...) be accurately extrapolated to location X (Y,Z,...)?

• How do we evaluate success?

• What is the the best estimator?
Introduction

our goal

• To build up a set of empirical case studies or data sandboxes on empirical validity...

...by finding experiments replicated in a sufficient number of times and places
... and using them in an empirical exploration of aspects of external validity.

• An empirically inductive epistemology.
Introduction

our first effort

• Apply the Angrist-Evans (1998) (same sex of the first two children increases the probability of further children and reduces the probability of the mother working) to International IPUMS data on 166 country-years.

• To develop a framework and present initial results on external validity.
In 2010, our sample of countries represents about 5 billion people (world population ≈ 7 billion).

Our working sample ≈ 550 million (women, aged 21-35, ≥2 kids)
Introduction

±

• Disadvantages.
  – Not a policy experiment.
  – Not an RCT.
  – Represents country-year level variation (for now).
  – External validity of external validity.

• But with the leap of faith some advantages.
  – Nationally representative data.
  – Wide geographical and temporal variation.
  ➤ Can address empirically a wide range of questions on external validity.
Question 1

• How do differences between reference and target locations drive extrapolation error.
  – We call this the extrapolation error (or X) function.
Question 1

set-up

- $T = 0, 1$ (control vs treatment of interest)
- Potential outcomes $Y(1), Y(0)$
- $D$
  - $= 0$, settings where we have experimental “reference settings”
  - $= 1$, settings where we wish to extrapolate, “target settings”
- A valid experiment:
  - Can be replaced by an observational study
Question 1

assumptions

(C1) Unconfounded location: that we have the information needed to characterize relevant differences between a reference and target contexts
  – Includes both micro covariates and experiment-level context variables

(C2) Common support: the range of values of covariates that characterize the target context are also observed reference context:
  – Particularly challenging for context characteristics

(C3) Functional form: that our estimator is sufficiently flexible to capture any heterogenous treatment effects
• Assume (C1), (C2), and (C3).
  – (C3) is readily satisfied by standard non-parametric estimators or by accepting linear approximations to non-linear functions.
  – (C2) is readily satisfied in our data, but testable.
  – We assume (C1) and then test for the validity of the assumption as described next.
Question 1

*how we extrapolate*

Compute the treatment effect in the reference location conditional on covariate values.

Use those estimates in the target location.

Reweight based on the covariate distribution in the target.
Question 1

the extrapolation error function

• Define the extrapolation error function as:
  • $X(\text{target covariates, reference covariates}) =$
    Extrapolated treatment effect
    – actual treatment effect

• If target covariates = reference covariates and $X \neq 0$, then (C1) has failed.

• For target covariates $\neq$ reference covariates, $X()$ maps out the extrapolation error function and violation of (C2), common support
Question 1

\textit{x-function estimation: dyadic approach}

- Create all country-year pairs (A,X);
- Use A to predict treatment effect in X (using covariate reweighting);
- Compare to actual treatment effect in X to estimate extrapolation error;
- Note covariate differences between A and X.
Question 1

the estimated $\varepsilon$ function

• Dyadic regressions (cont’d)
  – Given the set of bias vectors estimate a multivariate regression of the form:
    
    $$Extrapolation\ error = a + b\Delta W + e$$

    where regressions are weighted to accounted for estimated first stage and standard errors are clustered at the target country level.
  – Gives conditional extrapolation error function: the effect of one element of $W$ conditional on others.
Average effect = 0.04
~ US effect 0.06
~ mean more kids 0.57
≈ 6% effect

Funnel plot: ATE vs se.
Expect 95% of points to be within standard error lines, if effect homogenous.

Cochran’s Q-test and weighted Shapiro-Francia tests reject constant treatment at any standard level.
Question 1

$X$ function & education differences

- Treatment effect = 0.04
- $1 \sigma = 1$ on 4 pt ed scale
- $\epsilon = 0.08$

Test of unconfounded location assumption
Question 1

**GDP per capita differences**

Treatment effect = 0.04

1 σ = $10k

\[ \epsilon = 0.09 \]
Question 1

**LFP differences**

Treatment effect = 0.04

\[ 1 \sigma = 0.2 \text{ LFP} \]

\[ \epsilon = 0.2 \]

Mean prediction error +/- 2 se

-4 -2 0 2 4

Difference in LFP between target country and comparisons

Mean prediction error +/- 2 se
Question 1
within-region geographical distance

Treatment effect = 0.04
1 \( \sigma = 4800 \text{ km} \)
→ more than 10k km before significant effect
Question 1

temporal distance

Treatment effect = 0.04

\[ 1 \sigma = 10 \text{ years} \]

\[ \varepsilon = 0.04 \]
Nonparametric Lasso

Appendix Figure 4: LASSO solution paths for series approximation interaction terms

Panel A: Full solution path for "more kids" interaction terms

Panel B: Full solution path for "economically active" interaction terms

Panel C: Zoom view of solution path for "more kids" interaction terms

Panel D: Zoom view of solution path for "economically active" interaction terms
Table 4: Lasso regression for *Having more children*

<table>
<thead>
<tr>
<th>Step</th>
<th>Cp</th>
<th>R-squared</th>
<th>Level of variable</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4625.279</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2498.7156</td>
<td>0.0003</td>
<td>Country-Year</td>
<td>log GDP per capita</td>
</tr>
<tr>
<td>3</td>
<td>2350.8823</td>
<td>0.0003</td>
<td>Country-Year</td>
<td>Age mother</td>
</tr>
<tr>
<td>4</td>
<td>1584.9652</td>
<td>0.0004</td>
<td>Country-Year</td>
<td>Education mother level3</td>
</tr>
<tr>
<td>5</td>
<td>1376.3483</td>
<td>0.0004</td>
<td>Individual</td>
<td>Education mother level3</td>
</tr>
<tr>
<td>6</td>
<td>1363.2034</td>
<td>0.0004</td>
<td>Country</td>
<td>Region 5</td>
</tr>
<tr>
<td>7</td>
<td>1300.66</td>
<td>0.0004</td>
<td>Individual</td>
<td>Education spouse level3</td>
</tr>
<tr>
<td>8</td>
<td>982.5631</td>
<td>0.0004</td>
<td>Individual</td>
<td>Education mother level2</td>
</tr>
<tr>
<td>9</td>
<td>984.0331</td>
<td>0.0004</td>
<td>Country</td>
<td>Region 6</td>
</tr>
<tr>
<td>10</td>
<td>880.924</td>
<td>0.0005</td>
<td>Country-Year</td>
<td>Age mother second born</td>
</tr>
<tr>
<td>11</td>
<td>878.0398</td>
<td>0.0005</td>
<td>Individual</td>
<td>Education spouse level4</td>
</tr>
<tr>
<td>12</td>
<td>834.6946</td>
<td>0.0005</td>
<td>Country</td>
<td>Region 4</td>
</tr>
<tr>
<td>13</td>
<td>787.947</td>
<td>0.0005</td>
<td>Individual</td>
<td>Age mother</td>
</tr>
<tr>
<td>14</td>
<td>788.1958</td>
<td>0.0005</td>
<td>Country</td>
<td>Decade 1980</td>
</tr>
<tr>
<td>15</td>
<td>731.6514</td>
<td>0.0005</td>
<td>Country</td>
<td>Region 7</td>
</tr>
<tr>
<td>16</td>
<td>701.0733</td>
<td>0.0005</td>
<td>Country-Year</td>
<td>Education spouse level3</td>
</tr>
<tr>
<td>17</td>
<td>681.3512</td>
<td>0.0005</td>
<td>Individual</td>
<td>Education spouse level2</td>
</tr>
<tr>
<td>18</td>
<td>541.2938</td>
<td>0.0005</td>
<td>Country</td>
<td>Decade 1970</td>
</tr>
<tr>
<td>19</td>
<td>489.2243</td>
<td>0.0005</td>
<td>Country</td>
<td>Decade 1990</td>
</tr>
</tbody>
</table>
Question 1

Conditional $\in$ function more kids
Question 1

Conditional $\epsilon$ function more kids

Difference in mean education between target country and comparisons

Unconditional bias +/- se
Conditional bias +/- se

Difference in GDP per capita between target country and comparisons

Unconditional bias +/- se
Conditional bias +/- se
Question 2

- Where should we run experiments?
Question 2
Where to locate experiments? In the middle

Figure 19: Mean prediction error on percentile of comparison country composite treatment-effect predictor, using one site to predict all others

Notes: On the x-axis, each country, year is ranked based on its percentile of a composite treatment effect predictor. The composite predictor is a weighted average of country, year covariates weighted by their effect on the country, year treatment effect. The y-axis shows the mean prediction error from using the site on the x-axis to predict all other country, years. Source: Authors’ calculations based on data from the Integrated Public Use Microdata Series, International (IPUMS, I).

Figure 20: Mean prediction error on average Mahalanobis distance of the comparison country-year to all target country-years

Notes: On the x-axis, each country, year is ranked based on its average Mahalanobis distance to all other country, years. The y-axis shows the mean prediction error from using the site on the x-axis to predict all other country, years. Source: Authors’ calculations based on data from the Integrated Public Use Microdata Series, International (IPUMS, I).

Figure 21: Mean prediction error, given the first comparison site, on percentile of composite treatment-effect predictor covariate, using two sites to predict the others

Notes: On the x-axis, each country, year is ranked based on its percentile of a composite treatment effect predictor. The composite predictor is a weighted average of country, year covariates weighted by their effect on the country, year treatment effect. The y-axis shows the mean prediction error from using the site on the x-axis to predict all other country, years. Source: Authors’ calculations based on data from the Integrated Public Use Microdata Series, International (IPUMS, I).

Figure 22: Mean prediction error, given the first comparison site, on average Mahalanobis distance of the comparison country-year to all target country-years, using two sites to predict others

Notes: On the x-axis, each country, year is ranked based on its average Mahalanobis distance to all other country, years. The y-axis shows the mean prediction error from using the site on the x-axis in addition to the first selected comparison site to predict all other country, years. Source: Authors’ calculations based on data from the Integrated Public Use Microdata Series, International (IPUMS, I).
Question 3

• Are individual (subject) or country-level (experiment) covariates more important? Some have argued that between-country variation will dominate issues of external validity, so that extrapolating across countries is doomed.
Question 3

more kids

Density estimate: extrapolation error

CDF: absolute extrapolation error

Panel A: Density estimate - prediction error

Panel B: CDF - absolute prediction error

Notes: The graph plots the density estimates of the prediction error and CDF of the absolute prediction error based on the procedure described in Section 9 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series International (IPUMSKI).
Question 4

economically active

Density estimate: extrapolation error

CDF: absolute extrapolation error

- Bias: economically active
- CDF: absolute bias: economically active
Question 4
rules of thumb

• If extrapolation is possible, how should we do it?
• Using the model?
• Or might rules of thumb suffice?
Out-of-sample accuracy: model vs rules-of-thumb for more kids

[Fixed target] [no sex selection]
Question 5

To experiment or extrapolate?

- Suppose a decision maker wants to make an evidence-based decision of whether or not to implement a treatment. The decision maker has a choice between using the existing evidence base versus generating new evidence by carrying out an experiment in the target context.
- Many loss functions possible.
- For now, assume that the decision maker will decide that the existing evidence is sufficient to determine policy if a 95% prediction interval surrounding the conditional mean prediction for the target site is entirely on one or another side of some critical threshold, $c^*$. 
To experiment or extrapolate?

\[ \text{Var}[\tau|X = X_0] = \text{Var}[\hat{\tau}(X)|X = X_0] + \text{Var}[\epsilon(X)|X = X_0]. \]

Notes: Solid line = experiment not warranted. Dashed line = experiment warranted.
To experiment or extrapolate?
Question 6

choice of estimators

• Work in progress!
• Evident that dyadic approach is not efficient.
• Matching more efficient, but computationally challenging.
• Machine learning is a new option.
• Our results suggestion that flexible methods tend to overfit context covariates.
Question 7

*external validity of external validity*

- Does any of this generalize to the situations?
- Expand the data set.
- Look at other policy experiments.
Question 7
external validity of external validity

429 country-years, stretching from 1787 to 2014
Conclusions

five prescriptions

• The reference and target setting must be similar ($< \frac{1}{2} \sigma$) along economically relevant dimensions.

• Experiment-level data proved key in extrapolating.

• A sufficiently large experimental evidence base is needed for reliable extrapolation.

• Accounting for treatment effect heterogeneity is essential in extrapolating the treatment effect.

• With sparse data, modeling treatment effect heterogeneity is important; in data-rich settings, rules of thumb might be sufficient.
The X-Team

Local to global: experiments
Rajeev Dehejia Kiki Pop-Eleches Cyrus Samii
(Radon) (K-Pop) (Cyborg)
with James Bisbee (Buzzsaw)
Local to global: IV
with Michael Gechter (Seismograph)
The extrapolation machine: honing estimators
with Dan Aaronson (Dano)
External validity in the long-run: labor supply from 1800-2010

with Miikka Rokkanen (The Rok)
and Miguel Urquiola (MiG)
External validity in RDD