FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

IAPRI-MSU Technical Training

Impact Evaluation: Introduction & Methods Overview

Facilitated by Nicole M. Mason & Hambulo Ngoma

13 July 2017
Indaba Agricultural Policy Research Institute
Lusaka, Zambia

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

What is Impact Evaluation?

- "An impact evaluation **assesses changes in the well-being** of individuals, households, communities or firms **that can be attributed** to a particular project, program or policy" Source: World Bank
- "Impact evaluation is an **assessment of how the intervention** being evaluated **affects outcomes**, whether these effects are intended or unintended." Source: OECD
- "**The primary purpose of impact evaluation is to determine whether a program has an impact** (on a few key outcomes), and more specifically, to **quantify *how large* that impact is**." Source: J-PAL

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Motivating Examples

- What are some **examples** of project, programs, or policies for which you have conducted or might want to conduct an impact evaluation?
- Which **method(s)** did you use and why?

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

- "The **key challenge** in impact evaluation is finding a group of people who did not participate, but closely resemble the participants had those participants not received the program. Measuring outcomes in this **comparison group** is as close as we can get to measuring 'how participants would have been otherwise'. **There are many methods of doing this and each method comes with its own assumptions**." J-PAL Introduction to Evaluations

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

TODAY'S SESSION

1. Definition of impact evaluation (IE)
2. Motivating examples
3. **Key challenges in IE**
   - The missing counterfactual
   - Selection bias
   - External validity and internal validity
4. **Overview of common methods for IE**
   - Randomized evaluation
   - Propensity Score Matching
   - Difference-in-Differences
   - Instrumental Variables
   - Regression Discontinuity

FUTURE SESSIONS: GO INTO EACH METHOD IN DETAIL

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

By the end of today's session, you should be able to:

1. Define ***impact evaluation*** (IE)
2. Define ***counterfactual*** and explain why it's key to IE but not observable
3. Identify and explain 2 ***"counterfeit counterfactuals"*** (Khandker et al. 2009)
4. Define ***selection bias*** and explain why it's a problem
5. Explain the ***intuition of the various IE methods*** we cover, how they try to address the missing counterfactual and selection bias problems, and some of their main assumptions
6. Define ***external validity*** and ***internal validity***

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

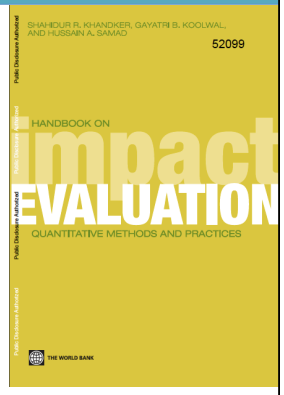
Main Reference

These training materials draw heavily on:

Khandker, S.R., Koolwal, G.B. and Samad, H.A., 2009. *Handbook on impact evaluation: quantitative methods and practices*. Washington, DC: World Bank Publications.

Available here.

We will not do Stata exercises today b/c this is an overview session, but there are Stata exercises on each method at the end of the Handbook.



USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

We will focus on quantitative methods for ex post IE

But there are various types of IE:

- **Ex ante** & **ex post**
- **Quantitative** & **qualitative**

Following Khandker et al. (2009)
Following Khandker et al. (2009)

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

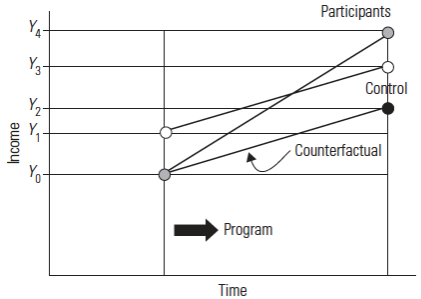
The missing counterfactual problem

- Suppose we want to do an IE of the Food Security Pack Program (FSPP) and are interested in how the program affected HH per capita income
- Consider a HH that participated in FSPP
  - *What outcome do we observe & what is the counterfactual?*
- Now consider a HH that did NOT participate
  - *Now what do we observe & what is the counterfactual?*
- The impact of FSPP can be measured by comparing observed and counterfactual HH per capita income
  - *What is the challenge/problem?*

WITH VS. WITHOUT

- The key comparison we want to make in IE is between outcomes **WITH VS. WITHOUT** the intervention (project/program/policy)
- **Impact = "With" outcome – "without" outcome**

Program Impact: With vs. Without



Source: Khandker et al. (2009)

Participants' income WITH the program?
- Y_4

Participants' income WITHOUT the program (counterfactual income)?
- Y_2

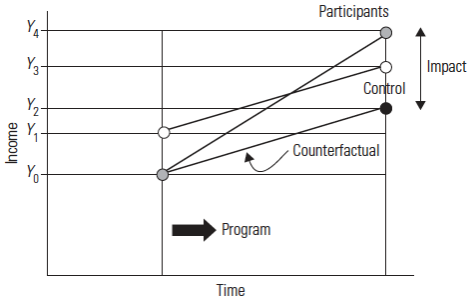
Program impact?
- $Y_4 - Y_2$

"Counterfeit counterfactuals" (Khandker et al. 2009)

Counterfeit counterfactual #1:

*Why can't we (in most cases) get a good IE estimate simply by comparing average outcomes after the intervention of **participants vs. non-participants**?*

Counterfeit counterfactual #1: Participants vs. Non-Participants



Source: Khandker et al. (2009)

Non-participants = "control" here

What is the (counterfeit counterfactual) impact if compare participants and non-participants?

- $Y_4 - Y_3$

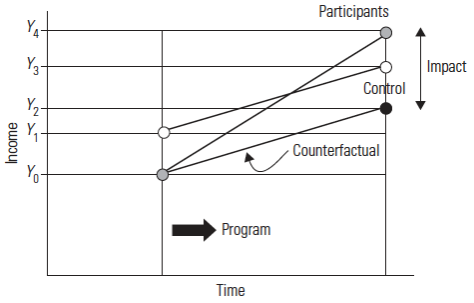
How does this compare to the true impact ($Y_4 - Y_2$) and why? (Hint: look at participants' and non-participants' incomes BEFORE the program.)

"Counterfeit counterfactuals" (Khandker et al. 2009)

Counterfeit counterfactual #2:

*Why can't we (in most cases) get a good IE estimate simply by comparing outcomes of participants **before vs. after** the intervention?*

Counterfeit counterfactual #2: Before vs. After



Source: Khandker et al. (2009)

What is the (counterfeit counterfactual) impact if compare participants' outcomes before vs. after the program?

- $Y_4 - Y_0$

How does this compare to the true impact ($Y_4 - Y_2$) and why? (Hint: what happens to the counterfactual over time?)

IE is about dealing with the missing counterfactual

Per Khandker et al. (2009, p. 25):

- "An impact evaluation is essentially a **problem of missing data**, because one cannot observe the outcomes of program participants had they not been beneficiaries."
- "Without information on the counterfactual, the **next best alternative is to compare outcomes of treated** individuals or households with those of a **comparison group** that has not been treated."
- The key is to "**pick a comparison group that is very similar to the treated group**, such that those who received treatment would have had outcomes similar to those in the comparison group in absence of treatment."

Selection bias

Notation follows Khandker et al. (2009)

Why not just estimate the following equation to obtain an unbiased estimate of the impact of the intervention?

$$Y_i = \beta T_i + \mathbf{X_i}\boldsymbol{\alpha} + \varepsilon_i$$

Where:

T_i = 1 if participate, =0 o.w.

$\mathbf{X_i}$ = vector of observed factors affecting Y

ε_i = error term (unobserved factors affecting Y)

Implications for OLS if Cov(T, ε)≠0?

Notation follows Khandker et al. (2009)

Why might treatment be non-random?

Following Khandker et al. (2009)

1. **Program placement**
2. **Self-selection**

Can be based on observed and/or unobserved factors

- *Which is more problematic – unobserved or observed?*
- **If based on unobserved factors**, then Cov(T, ε)≠0 and have (unobserved) **selection bias** → biased estimates

More formal notation: The potential outcomes framework (a.k.a. the Rubin causal model)

Notation follows Khandker et al. (2009)

- Let $Y_i(1)$ = PC income for HH i with treatment (T_i=1)
- Let $Y_i(0)$ = PC income for HH i without treatment (T_i=0)

- Impact for HH i = $Y_i(1)$-$Y_i(0)$ but counterfactual is unobserved
- In the population, the **average treatment effect** is:

  ATE = $E[Y_i(1)-Y_i(0)]$

Notation follows Khandker et al. (2009)

More formal notation (cont'd)

- Suppose we try to measure the ATE by comparing the average outcomes of participants and non-participants:

  $$\mathbf{D=E[Y_i(1)|T_i=1]-E[Y_i(0)|T_i=0]}$$

- *Does this equal the ATE $\{E[Y_i(1)-Y_i(0)]\}$?*
  - Only if $E[Y_i(1)|T_i=1]=E[Y_i(1)]$ and $E[Y_i(0)|T_i=0]=E[Y_i(0)]$
- To see selection bias, add/subtract $E[Y_i(0)|T_i=1]$ (the counterfactual):

  $D=E[Y_i(1)|T_i=1]-E[Y_i(0)|T_i=0]$**$+E[Y_i(0)|T_i=1]-E[Y_i(0)|T_i=1]$**
- Rearrange: $D=\underbrace{E[Y_i(1)|T_i=1]-E[Y_i(0)|T_i=1]}_{\text{ATT}}+\underbrace{E[Y_i(0)|T_i=1]-E[Y_i(0)|T_i=0]}_{\text{Selection bias}}$
- **If no selection bias, D=ATT** (can also write as $E[Y_i(1)-Y_i(0)|T_i=1]$
- **ATT=ATE if the potential outcomes are independent of treatment**

IE methods are different approaches to eliminate or correct for selection bias

In order to obtain unbiased estimates of the **causal effects** of the intervention

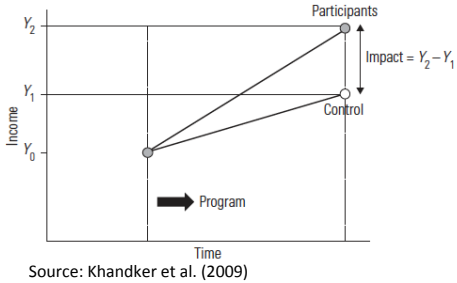
Notation follows Khandker et al. (2009)

Randomized evaluations

- *How do randomized evaluations deal with the selection bias problem?*
  - Randomize treatment
  - If done properly, then treatment assignment is independent of subject's characteristics
  - $D=\underbrace{E[Y_i(1)|T_i=1]-E[Y_i(0)|T_i=1]}_{\text{ATT}}+\underbrace{E[Y_i(0)|T_i=1]-E[Y_i(0)|T_i=0]}_{\text{Selection bias}}$
  - As a result, conditional=unconditional expectation, resulting in D=ATT, and ATT=ATE

Randomized evaluations (cont'd)

The Ideal Experiment with an Equivalent Control Group



Participants and control group similar before the program.

How can we see this in the figure?

Source: Khandker et al. (2009)

Notation follows Khandker et al. (2009)

Randomized evaluations (cont'd)

If randomization is at the i level (and some other assumptions hold), then estimate treatment effect via simple OLS:

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

where $Y_i = [Y_i(1) \bullet T_i] + [Y_i(0) \bullet (1-T_i)]$ is the observed outcome

Thought exercise: What randomized evaluation would you like to carry out if anything were possible?

Per Khandker et al. (2009)

There are many complexities to randomized evaluations (future session?) and some concerns – for example:

1. Ethical concerns
2. Compliance issues
3. Spillover effects
4. External validity

Aside: Internal vs. external validity

- **Validity**: "whether a particular conclusion or inference represents a good approximation to the true conclusion or inference (i.e., whether our methods of research and subsequent observations provide an adequate reflection of the truth)"
- **Internal validity**: "the ability of a researcher to argue that observed correlations are causal"
- **External validity**: "the ability to generalize the relationships found in a study to other persons, times, and settings"

Source: Roe & Just (2009, p. 1266)

Internal vs. external validity (cont'd)

	Relative Internal Validity	Relative External Validity	Topic and Subject Limits	Replicable?
Lab Experiments	High	Low	Long duration topics, larger stakes, losses	High
Field Experiments	Medium to High	Medium to High	Limited by researcher connections	Low to medium
Natural Experiments	Medium to High	High	Limited by occurrences of nature and policy	Low
Field/market Data	Low	High	Limited by privacy, recall and trade secrets	Low to medium

Figure 1. Tradeoffs across research methodologies

Source: Roe & Just (2009)

For what types of data/ research methodologies is internal validity more of a challenge and why?

How about external validity?

Implications?

Propensity Score Matching (PSM)

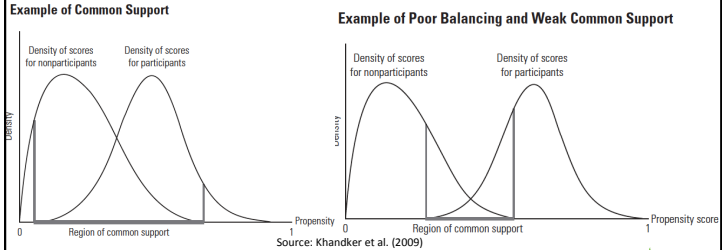
- *Who has used PSM before and what were you studying?*
- *What is the PSM approach to constructing a comparison group / approximating the counterfactual, and how is the ATE calculated?*
  - "PSM constructs a statistical **comparison group** that is **based on** a model of the **probability of participating** in the treatment, **using observed characteristics**. Participants are then **matched on the basis of this probability, or *propensity score***, to non-participants. The **average treatment effect** of the program is then calculated as the **mean difference in outcomes across the two groups**" (Khandker et al. 2009, p. 53)

Propensity Score Matching (PSM) – cont'd

- A critical PSM assumption is **selection on observables**. *What do you think this means?*
  - Other names for this assumption are **unconfoundedness**, **ignorability of treatment**, and **conditional independence** – i.e., conditional on observed covariates, treatment status is independent of the potential outcomes
  - If assume only $Y_1(0)$ is conditionally independent (weaker, less restrictive assumption), then get ATT instead of ATE
- *Implications for the circumstances under which PSM "solves" the selection bias problem?*
  - "**when only observed characteristics…affect program participation**" (Khandker et al., 2009 – p. 53)

Propensity Score Matching (PSM) – cont'd

- The other critical PSM assumption is that there is sufficient **overlap** (a.k.a. "**common support**") in the propensity scores of ultimate participants and non-participants



Source: Khandker et al. (2009)

Propensity Score Matching (PSM) – cont'd

- Many more PSM issues, intricacies, and related methods to discuss but hopefully this gives you the "gist" of the approach

Paraphrased from Khandker et al. (2009)

Difference-in-Differences (DID)

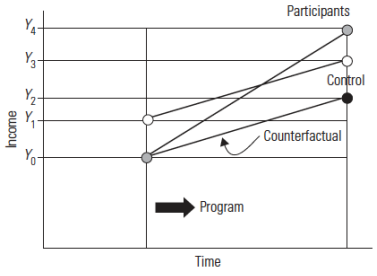
- **Key difference between PSM and DID**: **PSM** assumes **selection on observables only**, **DID** allows selection to be a function of **time-constant unobserved factors** (a.k.a. time invariant unobserved heterogeneity)
  - *Where have you heard this term before?*
  - *What if selection is a function of time-varying unobservables?*
- **Another key difference**:
  - **Randomized evaluations & PSM** – cross-sectional data sufficient (although panel data better – baseline/endline)
  - **DID requires panel data** (or at least repeated cross sections)

Difference-in-Differences (DID) – cont'd

- *Who has used DID before and what were you studying?*
- *What is the DID approach to constructing a comparison group / approximating the counterfactual, and how is the DID treatment effect calculated?*
  - "The DID estimator relies on a comparison of participants and non-participants before and after the intervention" (Khandker et al. 2009, p. 72)
  - DID Impact=(avg. ΔY participants)-(avg. ΔY non-participants)
    - **$(Y^T$ after- Y^T before) - $(Y^c$ after- Y^c before)**
    - → why it's called difference-in-differences or double difference

Difference-in-Differences (DID) – cont'd

DID impact = $(Y^T$ after- Y^T before) - $(Y^c$ after- Y^c before)



Source: Khandker et al. (2009)

Change in participants' income?
- Y_4-Y_0

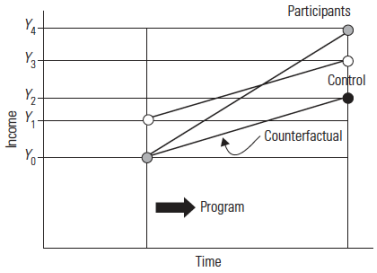
Change in non-participants' (control) income?
- Y_3-Y_1

DID impact?
- $(Y_4$-$Y_0)$-$(Y_3$-$Y_1)$

Difference-in-Differences (DID) – cont'd

Key assumption: parallel trends = "unobserved characteristics affecting program participation do not vary over time with treatment status" (Khandker et al. 2009, p. 73)

- Implies $(Y_1-Y_0)=(Y_3-Y_2)$



Source: Khandker et al. (2009)

Change in participants' income?

- Y_4-Y_0

Change in non-participants' (control) income?

- Y_3-Y_1

DID impact?

- $(Y_4-Y_0)-(Y_3-Y_1)$
- $=Y_4-Y_0-Y_3+Y_1=Y_4-Y_3+Y_1-Y_0$

Substitute in $(Y_1-Y_0)=(Y_3-Y_2)$ (parallel trend assumption):

- $= Y_4-Y_3+Y_3-Y_2= Y_4-Y_2$

Same as with vs. without!

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Difference-in-Differences (DID) – cont'd

DID impact = (Y^T after- Y^T before) - (Y^c after- Y^c before)

The following table gives mean income during the pre- and postintervention period for a microfinance intervention in the rural Lao People's Democratic Republic:

	Mean income (KN thousand)	
	Participants	*Nonparticipants*
Preintervention period	80	90
Postintervention period	125	120

Impact of microfinance intervention on participants' income using DD is

(a) KN 45,000
(b) KN 30,000
(c) KN 15,000

Source: Khandker et al. (2009)

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Notation follows Khandker et al. (2009)

Difference-in-Differences (DID) – cont'd

Regression set-up with panel data (without control variables) where i indexes the individual or HH, and t indexes time, with t=1 after the program and t=0 before the program):

$$Y_{it} = \alpha + \rho T_{i1} + \gamma t + \beta T_{i1} t + \varepsilon_{it}$$

- *Which parameter is the DID impact estimate?*
  - β (parameter on the treatment*after term)

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Notation follows Khandker et al. (2009)

Difference-in-Differences (DID) – cont'd

Panel fixed effects model set-up (WITH control variables):

$$Y_{it} = \phi T_{it} + \mathbf{X_{it}}\boldsymbol{\delta} + \eta_i + u_{it}$$

- First difference to remove η_i:

$$\Delta Y_{it} = \phi \Delta T_{it} + \Delta \mathbf{x_{it}}\boldsymbol{\delta} + \Delta u_{it}$$

- *Which parameter is the DID impact estimate?*
  - ϕ

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

PSM - DID

- If have data on participants and non-participants before and after the program, then can combine PSM and DID
- **PSM – DID ATT**: difference in mean *changes* in outcomes (before vs. after the program) between participants and matched non-participants

Instrumental Variables (IV)

- Probably the method that you are most familiar with
- Covered in IAPRI training in May 2013; recently did similar training in Kenya → will send materials
- If have a valid IV, then IV approach can correct for time-varying selection bias (unlike PSM and DID)
- If combine with panel data, then can do FE-IV to address time-invariant and time-varying selection bias
- *So what 2 conditions must a candidate IV satisfy to be a valid IV?*

Instrumental Variables (IV) – cont'd

- **Two conditions for an IV to be valid:**
  1. **Strongly partially correlated with the endogenous explanatory variable** (1st stage partial F-stat > 10)
  2. **Uncorrelated with unobserved factors that affect the outcome** variable of interest
- See IV ppt slides for details
- Downside: very difficult to find valid IVs
- Related method: control function (CF) approach
  - Useful when using non-linear-in-parameters models (probit, Tobit, etc.)

ASIDE: Endogenous Switching Regression Models

- Hambulo will lead us through this discussion

Endogenous Switching Regression (ESR) Models

- Useful to study welfare effects of technology adoption, e.g.
  - self-selection can confound outcomes (think about only the best farmers selecting themselves to adopt technology X and you want to assess the impacts of adopting X on Y)
  - self-selection can cause endogeneity bias
- Because the reasons for selection may be systematic, selection and outcomes are correlated
- ESR models parcels observation units into two regimes (with one regime observed and the other unobserved). Unlike
  - Heckman set-up, ESR allows you to use the full sample
  - 2SLS and double hurdle, ESR allows you to get estimates for both adopters and non-adopters. These are needed to compute various impact assessment measures, e.g., ATT, ATU, ATE
- Identification requires exclusion restrictions (similar to an IV); need a variable in the selection equation not in the outcome equations

ESR set-up (brief)

- First Stage: define selection over a criterion function *I*

$$I_i = 1 \quad \text{if} \quad \gamma Z_i + u_i > 0$$
$$I_i = 0 \quad \text{if} \quad \gamma Z_i + u_i \leq 0$$

- Second Stage: two outcomes equations define the regimes

$$\text{Regime1}: \; y_{1i} = \beta_1 X_{1i} + \epsilon_{1i} \quad \text{if } I_i = 1$$
$$\text{Regime2}: \; y_{2i} = \beta_2 X_{2i} + \epsilon_{2i} \quad \text{if } I_i = 0$$

- Self-selection makes *corr (ui,e1,e2) ≠ 0. Regime specific inverse mills ratios needed in outcome equations*
- The two-steps can be estimated manually with OLS or MLE (need to correct standard errors) or using FIML *movestay in Stata* (more later, see refs for applications)
- Use conditional expectations to compute counterfactual outcomes

Some ESR Model References

- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.
- Lokshin, M., & Sajaia, Z. (2004). Maximum likelihood estimation of endogenous switching regression models. *Stata Journal, 4*(3), 282-289.
- Alem, Y., Eggert, H., & Ruhinduka, R. (2015). Improving Welfare Through Climate-Friendly Agriculture: The Case of the System of Rice Intensification. *Environmental and Resource Economics, 62*(2), 243-263. doi:10.1007/s10640-015-9962-5
- Abdulai, A., & Huffman, W. (2014). The adoption and impact of soil and water conservation technology: An endogenous switching regression application. *Land economics, 90*(1), 26-43.

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Regression Discontinuity (RD)

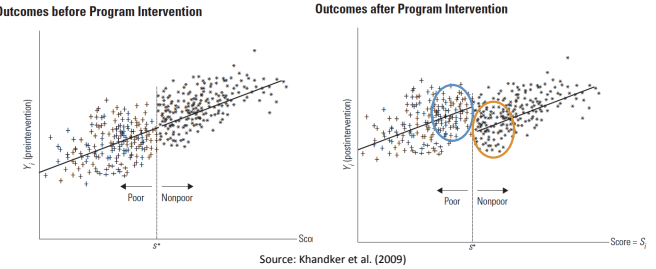
- *Who has used RD before and what were you studying?*
- **RD**: "**program eligibility rules** can sometimes be used as **instruments for exogenously identifying program participants and nonparticipants**. To establish comparability, one can **use participants and nonparticipants within a certain neighborhood of the eligibility threshold** as the relevant sample for estimating the treatment impact. Known as ***regression discontinuity (RD)***, this method allows **observed as well as unobserved heterogeneity** to be accounted for." (Khandker et al. 2009, p. 103)

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI

Regression Discontinuity (RD) – cont'd

- Similar to IV "because they introduce an exogenous variable that is highly correlated with participation, albeit not akin to participation" (Khandker et al. 2009, p. 104)
- Examples (from Khandker et al. 2009):
  - Grameen Bank program: HH landholding < 0.5 acre
  - Pension programs: eligible if above a specific age
- Zambia examples?
- RD challenges/concerns (per Khandker et al. 2009, p. 103):
  1. "Eligibility rules will not be adhered to consistently"
  2. "Potential for eligibility rules to change over time"

Regression Discontinuity (RD) – cont'd



Source: Khandker et al. (2009)

- Suppose poor HHs eligible for program, non-poor HHs not eligible
- **RD gist**: estimate the treatment effect by comparing the average outcomes of HHs/individuals just to the left vs. just to the right of the threshold (s^*)

We'd like to do future sessions dedicated to each of these methods.

- IV and intro to CF – Nicky will send ppt slides and Stata code from May 2017 training at Tegemeo
- Sept 2017 – Hambulo – ESR models
- Late 2017/early 2018 – possible session(s) on RCTs
- 2018 – possible sessions on PSM, DID, and RD

WHICH METHODS ARE MOST CRITICAL FOR US TO FOCUS ON GOING FORWARD?

Thank you for your attention & participation!

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Acknowledgements

This training was supported by funding from the U.S. Agency for International Development through the Zambia Mission Buy-In to the Innovation Lab for Food Security Policy.

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | IAPRI