Economics 174/274
Global Poverty and Impact Evaluation

Professor Frederico Finan

February 24, 2011

Lecture 9
Social scientists and policymakers traditionally thought of health purely as a measure of well-being, living standards

Clearly, two of the objectives of development are:
- to improve the material well-being of individuals, and
- to enjoy this material well-being during longer life-spans

Example: health/longevity components of human development index
- Life expectancy
- Infant/child survival rates
Health and Economic Development

- Alternative/complementary view: health as human capital

- Since the 1960s, however, economists have also started thinking of health as a form of human capital. If there is a return to the health of an individual, then investments in the health of the population may have important effects on economic development.

- In summary, individual health, as a form of human capital, may have important effects on development.
Health and Economic Development

Key topics for this module

▶ The global disease burden

▶ Health as human capital - productivity-enhancing

▶ The impact of the HIV/AIDS epidemic in LDCs
Health and Economic Development

Definition of health according to the World Health Organization [WHO]:

“A state of complete physical, mental, and social well-being and not merely the absence of disease and infirmity.”

- Despite outstanding achievements, the developing world continues to face great challenges as it seeks to continue to improve the health and education of its people

- Example: Child mortality rates in LDCs remain more than ten times higher than those found in rich countries

- These deaths generally result from conditions that are easily treatable (e.g., death from dehydration due to diarrhea)
Health and Economic Development

▶ Another measure of longevity: *life expectancy at birth*

▶ Note: these are coarse measures of health status of the population
  ▶ Extension of life can provide extended years of vitality in one country while providing only additional years of extremely poor health or suffering in another.

▶ Child (0-5 years) survival rates - do not measure health status of the general population past early childhood (although serves as a reasonable proxy)
The Global Disease Burden

- Major health problems in developing countries:
  - Diarrhea
  - Childhood diseases, including malaria, acute respiratory infections, parasitic worm infections, measles
  - Malnutrition (disease? or condition?)
  - HIV/AIDS

- These conditions (excluding HIV/AIDS) account for 70% of deaths among children less than five (5) years of age
- These health problems are particularly severe in Sub-Saharan Africa
  - Infant mortality > 100 per 1,000 births in many countries
The Global Disease Burden

- Major health problems in developing countries: Diarrhea

- 4 million children under the age of 5 die each year from diarrhea

- Death due to dehydration

- Source(s): contaminated water (e.g. fecal matter, other)

- Deaths concentrated in early childhood (0-5 years)
The Global Disease Burden

- **Treatment**
  - Promotion of breastfeeding
  - Oral rehydration therapy

- **Prevention**
  - Access to (abundant) clean water (rural areas - protect springs)
The Global Disease Burden: Diarrhea

- Jalan and Ravallion examine whether access to piped water reduces diarrhea for children in rural India

- They use a large, representative cross-sectional survey for rural India implemented in 1993-1994

- An estimated 1.5 million child deaths per year in India due to diarrhea and other diseases related to poor water quality

- 1/5 of the population of rural India do not have access to safe drinking water (World Bank, 2000)

- Expanding access to piped water is considered an important development action in India
Main question: Is a child less vulnerable to diarrhoeal disease if he/she lives in a household with access to piped water?
Main question: Is a child less vulnerable to diarrhoeal disease if he/she lives in a household with access to piped water?

How do we answer this question empirically?
Main question: Is a child less vulnerable to diarrhoeal disease if he/she lives in a household with access to piped water?

How do we answer this question empirically?

Can we simply compare households with and without piped water? Why not?
Main question: Is a child less vulnerable to diarrhoeal disease if he/she lives in a household with access to piped water?

How do we answer this question empirically?
Can we simply compare households with and without piped water? Why not?

Without random assignment, Jalan and Ravallion rely on propensity score matching.
What is Propensity-Score Matching (PSM)?

- When a treatment cannot be randomized, the next best thing to do is to try to mimic randomization.

- With matching methods, one tries to develop a counterfactual or control group that is as similar to the treatment group as possible in terms of observed characteristics.
What is Propensity-Score Matching (PSM)?

- We need to find a large group of nonparticipants who are observationally similar to participants in terms of characteristics not affected by the program.

- Each participant is matched with an observationally similar nonparticipant, and then the average difference in outcomes across the two groups is compared to get the program treatment effect.

- If one assumes that differences in participation are based solely on differences in observed characteristics, and if enough nonparticipants are available to match with participants, the corresponding treatment effect can be measured even if treatment is not random.
What is Propensity-Score Matching (PSM)?

► Issue: in practice, this can be quite hard
  ▶ There may be many important characteristics!
  ▶ A lot of bins - curse of dimensionality
  ▶ Hard to find two individuals who match along all the characteristics

► It would be useful to create an “index” that aggregates all these characteristics
► Rosenbaum and Rubin proposed a solution
  ▶ Compute everyone’s probability (or propensity) of participating, based on their observable characteristics
  ▶ Choose matches that have the same probability of participation as the treatments
What does PSM do?

- PSM constructs a statistical comparison group by modeling the probability of participating in the program on the basis of observed characteristics unaffected by the program.

- Participants are then matched on the basis of this probability, or propensity score, to nonparticipants.

- The average treatment effect of the program is then calculated as the mean difference in outcomes across these two groups.

- PSM is useful when only observed characteristics are believed to affect program participation.

- That is, there is no selection bias based on unobservable characteristics.
PSM Method in Theory

- PSM approach tries to capture the effects of different observed controls \( X \) on participation in a single propensity score or index.

- In particular, we estimate the following model
  \[
  P(X) = Pr(T = 1|X)
  \]

- Rosenbaum and Rubin (1983) show that, under certain assumptions, matching on \( P(X) \) is as good as matching on \( X \).
What are the necessary assumptions for a causal interpretation?

1. conditional independence
2. common support
Conditional independence

- **Conditional independence** states that given a set of observable covariates $X$ that are not affected by treatment, potential outcomes $Y$ are independent of treatment assignment $T$.

- If $Y_i^T$ represent outcomes for participants and $Y_i^C$ outcomes for nonparticipants, conditional independence implies

  $$(Y_i^T, Y_i^C) \perp T_i|X_i$$

- This assumption is also called *unconfoundedness*.
Conditional independence

- Is conditional independence a strong assumption?
Conditional independence

- Is conditional independence a strong assumption?
- When is violated?
Conditional independence

- Is conditional independence a strong assumption?
- When is violated?
- Is it testable?

Useful to have a lot of preprogram data (a lot of controls) and to understand the assignment mechanism.
Conditional independence

- Is conditional independence a strong assumption?
- When is violated?
- Is it testable?
- What assignment mechanism guarantees this assumption?
Conditional independence

► Is conditional independence a strong assumption?
► When is violated?
► Is it testable?
► What assignment mechanism guarantees this assumption?

► Useful to have a lot of preprogram data (a lot of controls) and to understand the assignment mechanism
Common Support

- This condition ensures that treatment observations have comparison observations “nearby” in the propensity score distribution

- $0 < P(T_i = 1|X_i) < 1$

- The effectiveness of PSM also depends on having a large and roughly equal number of participant and nonparticipant observations so that a substantial region of common support can be found
Example of Common Support

Density of scores for nonparticipants

Density of scores for participants

Density

Region of common support

Propensity score
Example of Poor Balance and Weak Common Support
Common support

- Treatment units will therefore have to be similar to non-treatment units in terms of observed characteristics unaffected by participation;
- Some non-treatment units may have to be dropped to ensure comparability.
- However, sometimes a nonrandom subset of the treatment sample may have to be dropped if similar comparison units do not exist.
- This may create possible sampling bias in the treatment effect.
- Examining the characteristics of dropped units may be useful in interpreting potential bias in the estimated treatment effects.
Estimating the treatment effect with PSM

1. Estimating a Model of Program Participation
2. Defining the Region of Common Support and Balancing Tests
3. Matching Participants to Nonparticipants
4. Computing the treatment effect
Estimating a Model of Program Participation

Estimate the probability of participation $T$ on a set of covariates $X$ that are likely to determine participation.
Estimating a Model of Program Participation

- Estimate the probability of participation \( T \) on a set of covariates \( X \) that are likely to determine participation.

- That is, we want to estimate \( Pr(T = 1|X) \). How do we do this?

Suppose for simplicity, that \( T_i = \beta_0 + \beta_1 X_i + u_i \)

- What is \( E(u_i|X) \)? and why?

\[ E(u_i|X) = 0 \] because of conditional independence
Estimating a Model of Program Participation

- Estimate the probability of participation \( T \) on a set of covariates \( X \) that are likely to determine participation.
- That is, we want to estimate \( Pr(T = 1 | X) \). How do we do this?
- Suppose for simplicity, that

\[
T_i = \beta_0 + \beta_1 X_i + u_i
\]
Estimating a Model of Program Participation

- Estimate the probability of participation $T$ on a set of covariates $X$ that are likely to determine participation.
- That is, we want to estimate $Pr(T = 1|X)$. How do we do this?
- Suppose for simplicity, that 

$$T_i = \beta_0 + \beta_1 X_i + u_i$$

- What is $E(u_i|X)$? and why?
Estimating a Model of Program Participation

- Estimate the probability of participation $T$ on a set of covariates $X$ that are likely to determine participation.
- That is, we want to estimate $\Pr(T = 1|X)$. How do we do this?
- Suppose for simplicity, that

$$T_i = \beta_0 + \beta_1 X_i + u_i$$

- What is $E(u_i|X)$? and why?
- $E(u_i|X) = 0$ because of conditional independence
Estimating a Model of Program Participation

- So given conditional independence we know that

\[ E[T_i | X_i] = \beta_0 + \beta_1 X_i \]

But what is the expected value of participation when \( T_i \) is binary?

\[ E[T_i | X_i] = 1 \times \Pr(T_i = 1 | X_i) + 0 \times \Pr(T_i = 0 | X_i) \]

\[ E[T_i | X_i] = \Pr(T_i = 1 | X_i) \]

\[ \Pr(T_i = 1 | X_i) = \beta_0 + \beta_1 X_i \]

But this is just the linear probability model, which we have already estimated! (OLS estimation when the dependent variable is binary)
So given conditional independence we know that

\[
E[T_i|X_i] = \beta_0 + \beta_1 X_i
\]

But what is the \(E[T_i|X_i]\) when \(T_i\) is binary?

\[
E[T_i|X_i] = \Pr(T_i = 1 | X_i) \times 1 + \Pr(T_i = 0 | X_i) \times 0
\]

\[
\Pr(T_i = 1 | X_i) = \beta_0 + \beta_1 X_i
\]

But this is just the linear probability model, which we have already estimated! (OLS estimation when the dependent variable is binary)
Estimating a Model of Program Participation

- So given conditional independence we know that
- \( E[T_i|X_i] = \beta_0 + \beta_1 X_i \)
- But what is the \( E[T_i|X_i] \) when \( T_i \) is binary?

\[
E[T_i|X_i] = \Pr(T_i = 1|X_i) \\
\Pr(T_i = 1|X_i) = \beta_0 + \beta_1 X_i
\]

But this is just the linear probability model, which we have already estimated! (OLS estimation when the dependent variable is binary)
So given conditional independence we know that

\[ E[T_i|X_i] = \beta_0 + \beta_1 X_i \]

But what is the \( E[T_i|X_i] \) when \( T_i \) is binary?

\[ E[T_i|X_i] = 1 \times Pr(T_i = 1|X_i) + 0 \times Pr(T_i = 0|X_i) \]
So given conditional independence we know that

\[ E[T_i|X_i] = \beta_0 + \beta_1 X_i \]

But what is the \[ E[T_i|X_i] \] when \( T_i \) is binary?

\[ E[T_i|X_i] = 1 \times Pr(T_i = 1|X_i) + 0 \times Pr(T_i = 0|X_i) \]

\[ E[T_i|X_i] = Pr(T_i = 1|X_i) \]
Estimating a Model of Program Participation

- So given conditional independence we know that
  \[ E[T_i|X_i] = \beta_0 + \beta_1 X_i \]

- But what is the \( E[T_i|X_i] \) when \( T_i \) is binary?
  \[ E[T_i|X_i] = 1 \times Pr(T_i = 1|X_i) + 0 \times Pr(T_i = 0|X_i) \]

- \( E[T_i|X_i] = Pr(T_i = 1|X_i) \)

- \( Pr(T_i = 1|X_i) = \beta_0 + \beta_1 X_i \)
Estimating a Model of Program Participation

- So given conditional independence we know that
  \[ E[T_i|X_i] = \beta_0 + \beta_1 X_i \]
- But what is the \( E[T_i|X_i] \) when \( T_i \) is binary?
  \[ E[T_i|X_i] = 1 \times Pr(T_i = 1|X_i) + 0 \times Pr(T_i = 0|X_i) \]
  \[ E[T_i|X_i] = Pr(T_i = 1|X_i) \]
  \[ Pr(T_i = 1|X_i) = \beta_0 + \beta_1 X_i \]
- But this is just the linear probability model, which we have already estimated! (OLS estimation when the dependent variable is binary)
Estimating a Model of Program Participation

Unfortunately, the linear probability model is not great for PSM. Why?
Estimating a Model of Program Participation

Unfortunately, the linear probability model is not great for PSM. Why?
The problem is the probability of $T = 1$ is modeled as linear: 

$$Pr(T_i = 1|X_i) = \beta_0 + \beta_1 X_i$$

We want instead: 

$$0 < Pr(T = 1|X) < 1$$

for all $X$ (Common support assumption)

This requires a nonlinear functional form for the probability
Two common options

1. Probit:

\[ Pr(T_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_i) \]

2. Logit:

\[ Pr(T_i = 1|X_i) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i)} \]
Estimating a Model of Program Participation
Estimating a Model of Program Participation

- Now that we have the functional form decided (assume a Logit)
Now that we have the functional form decided (assume a Logit)

What covariates should we include?

Typically we want to include variables that are correlated with program participation but not affected by the treatment.

We want the specification to be flexible (include interaction terms, higher polynomials)

We also want a specification that achieves a good degree of balance (thus we should be careful not to overfit the model)

This is not a behavioral model!
Now that we have the functional form decided (assume a Logit).

What covariates should we include?

Unfortunately there is not much guidance on how to select the X’s.

We want the specification to be flexible (include interaction terms, higher polynomials).

We also want a specification that achieves a good degree of balance (thus we should be careful to overfit the model).

This is not a behavioral model!
Estimating a Model of Program Participation

- Now that we have the functional form decided (assume a Logit)
- What covariates should we include?
- Unfortunately there is not much guidance on how to select the $X$'s
- Typically we want to include variables that are correlated with program participation but not affected by the treatment.
Estimating a Model of Program Participation

- Now that we have the functional form decided (assume a Logit)
- What covariates should we include?
- Unfortunately there not much guidance on how to select the $X$'s
- Typically we want to include variables that are correlated with program participation but not affect by the treatment.
- We want the specification to be flexible (include interaction terms, higher polynomials)
Now that we have the functional form decided (assume a Logit)

What covariates should we include?

Unfortunately there not much guidance on how to select the $X$’s

Typically we want to include variables that are correlated with program participation but not affect by the treatment.

We want the specification to be flexible (include interaction terms, higher polynomials)

We also want a specification that achieves a good degree of balance (thus we should be careful to overfit the model)
Estimating a Model of Program Participation

- Now that we have the functional form decided (assume a Logit)
- What covariates should we include?
- Unfortunately there not much guidance on how to select the X’s
- Typically we want to include variables that are correlated with program participation but not affect by the treatment.
- We want the specification to be flexible (include interaction terms, higher polynomials)
- We also want a specification that achieves a good degree of balance (thus we should be careful to overfit the model)
- This is not a behavioral model!
Defining the Region of Common Support and Balancing Tests

- Find the regions of distribution where the propensity score for treatment and control overlap

- Drop the observations where there is no common support

- Check the characteristics of non-participants that were dropped to see how much sampling bias may have occurred

- Conduct balancing tests within each quantile of the propensity score distribution. Make sure the $X$'s are balanced between treatment and control: $\hat{P}(X|T=1) = \hat{P}(X|T=0)$
Defining the Region of Common Support and Balancing Tests

- Find the regions of distribution where the propensity score for treatment and control overlap

- Drop the observations where there is no common support
Defining the Region of Common Support and Balancing Tests

- Find the regions of distribution where the propensity score for treatment and control overlap
- Drop the observations where there is no common support
- Check the characteristics of non-participants that were dropped to see how much sampling bias may have occurred
Defining the Region of Common Support and Balancing Tests

- Find the regions of distribution where the propensity score for treatment and control overlap

- Drop the observations where there is no common support

- Check the characteristics of non-participants that were dropped to see how much sampling bias may have occurred

- Conduct balancing tests within each quantile of the propensity score distribution. Make sure the $X$'s are balanced between treatment and control: $\hat{P}(X|T = 1) = \hat{P}(X|T = 0)$
Matching Participants to Nonparticipants

- Now that we have the propensity score $\hat{P}(X)$, how do we match participants to non-participants?
Matching Participants to Nonparticipants

Now that we have the propensity score $\hat{P}(X)$, how do we match participants to non-participants?

Do we only match those with the same probability? Or do match those whose probabilities are close? What does it mean to be close? We need a metric.
Matching Participants to Nonparticipants

- Now that we have the propensity score $\hat{P}(X)$, how do we match participants to non-participants?

- Do we only match those with the same probability? Or do we match those whose probabilities are close? What does it mean to be close? We need a metric.

- There are many different matching criteria one could use to assign participants to non-participants on the basis of the propensity score.
Matching Participants to Nonparticipants

- Nearest-neighbor matching

- Each treatment unit is matched to the comparison unit with the closest propensity score
- One can also choose $n$ nearest neighbors and do matching (usually $n = 5$ is used)
- Matching can be done with and without replacement
- Potential issue: 5 closest may still be far apart

- Caliper or radius matching

- Imposes a threshold or tolerance on the maximum propensity distance
- Potential issue: higher number of dropped nonparticipants is likely → increasing the chance of sampling bias
Matching Participants to Nonparticipants

- Nearest-neighbor matching
  - Each treatment unit is matched to the comparison unit with the closest propensity score
Matching Participants to Nonparticipants

- Nearest-neighbor matching
  - Each treatment unit is matched to the comparison unit with the closest propensity score
  - One can also choose n nearest neighbors and do matching (usually \( n = 5 \) is used)
Matching Participants to Nonparticipants

- Nearest-neighbor matching
  - Each treatment unit is matched to the comparison unit with the closest propensity score
  - One can also choose n nearest neighbors and do matching (usually n = 5 is used)
  - Matching can be done with and without replacement
Matching Participants to Nonparticipants

- Nearest-neighbor matching
  - Each treatment unit is matched to the comparison unit with the closest propensity score
  - One can also choose n nearest neighbors and do matching (usually n = 5 is used)
  - Matching can be done with and without replacement
  - Potential issue: 5 closest may still be far apart
Matching Participants to Nonparticipants

- Nearest-neighbor matching
  - Each treatment unit is matched to the comparison unit with the closest propensity score
  - One can also choose $n$ nearest neighbors and do matching (usually $n = 5$ is used)
  - Matching can be done with and without replacement
  - Potential issue: 5 closest may still be far apart

- Caliper or radius matching
Matching Participants to Nonparticipants

- Nearest-neighbor matching
  - Each treatment unit is matched to the comparison unit with the closest propensity score
  - One can also choose n nearest neighbors and do matching (usually n = 5 is used)
  - Matching can be done with and without replacement
  - Potential issue: 5 closest may still be far apart

- Caliper or radius matching
  - Imposes a threshold or tolerance on the maximum propensity distance
Matching Participants to Nonparticipants

▶ Nearest-neighbor matching
  ▶ Each treatment unit is matched to the comparison unit with the closest propensity score
  ▶ One can also choose \( n \) nearest neighbors and do matching (usually \( n = 5 \) is used)
  ▶ Matching can be done with and without replacement
  ▶ Potential issue: 5 closest may still be far apart

▶ Caliper or radius matching
  ▶ Imposes a threshold or tolerance on the maximum propensity distance
  ▶ Potential issue: higher number of dropped nonparticipants is likely \( \rightarrow \) increasing the chance of sampling bias
Matching Participants to Nonparticipants

- Stratification or interval matching

  - Partition data into different strata (usually quintiles) and calculate the program's impact with each interval (i.e. mean difference in outcomes between treated and control observations).
  - Overall impact is a weighted average of these interval impacts (weighted by share of participants).

- Kernel and local linear matching

  - Nonparametric technique uses a weighted average of all nonparticipants to construct the counterfactual match for each participant.
Matching Participants to Nonparticipants

- Stratification or interval matching
  - Partition data into different strata (usually quintiles) and calculate the program’s impact with each interval (i.e. mean difference in outcomes between treated and control observations).
Matching Participants to Nonparticipants

- Stratification or interval matching
  - Partition data into different strata (usually quintiles) and calculate the program’s impact with each interval (i.e. mean difference in outcomes between treated and control observations).
  - Overall impact is a weighted average of these interval impacts (weighted by share of participants).
Matching Participants to Nonparticipants

- Stratification or interval matching
  - Partition data into different strata (usually quintiles) and calculate the program’s impact with each interval (i.e. mean difference in outcomes between treated and control observations).
  - Overall impact is a weighted average of these interval impacts (weighted by share of participants)

- Kernel and local linear matching
Matching Participants to Nonparticipants

- **Stratification or interval matching**
  - Partition data into different strata (usually quintiles) and calculate the program’s impact with each interval (i.e. mean difference in outcomes between treated and control observations).
  - Overall impact is a weighted average of these interval impacts (weighted by share of participants)

- **Kernel and local linear matching**
  - Nonparametric technique uses a weighted average of all nonparticipants to construct the counterfactual match for each participant
Matching Participants to Nonparticipants

Regression methods

- Estimate the following regression:

\[ Y_i = \beta_0 + \beta_1 T_i + \beta_2 \hat{P}(X) + \beta_3 \left( T_1 \times (\hat{P}(X) - E[\hat{P}(X)]) \right) + u_i \]

- Estimate the following regression: \( Y_i = \beta_0 + \beta_1 T_i + \delta X_i + u_i \)
  using weighted least squares

- Weights are 1 for participants and \( \frac{\hat{P}(X)}{1-\hat{P}(X)} \) for non-participants
PSM

STATA EXAMPLE
Main question: Is a child less vulnerable to diarrhoeal disease if he/she lives in a household with access to piped water?

Do children in poor, or poorly educated, households realize the same health gains from piped water as others?
propensity-score matching (PSM) methods to estimate the causal effects of piped-water on child health in a cross-sectional sample without random placement

Two groups: those households that have piped water \( D_i = 1 \) and those that do not \( D_i = 0 \)

Radius and nearest neighbor matching: The nearest neighbor to the \( i^{th} \) participant is defined as the non-participant that minimizes \([p(x_i) - p(x_j)]^2\) over all \( j \) in the set of non-participants, then uses 5 nearest

\( p(x_k) \) is the predicted odds ratio of observation \( k \rightarrow p(x_k) = \frac{\hat{p}(x_k)}{1 - \hat{p}(x_k)}\)

Matches were accepted if \([p(x_i) - p(x_j)]^2 < 0.001\)
Estimate of the impact: Let $\Delta H_j$ denote the gain in health status for child $j$

$$\Delta H = \sum_{j=1}^{T} \omega_j \left( h_{j1} - \sum_{j=1}^{C} W_{ij} h_{ij0} \right)$$

- $\omega_j$ - sampling weights
- $W_{ij}$ - weights applied in calculating the average
- $h_{ij0}$ - outcome indicator of the $i^{th}$ non-treated matched to the $j^{th}$ treated
- $h_{j1}$ - post-intervention health indicator
Nationally representative survey collecting detailed information on education and health status of 33,000 rural households from 1765 villages covering 16 states of India

Access to piped water - an indicator for whether the household reports access to piped water from a tap either inside or outside the house

Outcome variable - prevalence of diarrhea among children under 5 years of age and the reported illness duration
## Summary stats

### Table 1
Access to piped water across the income distribution and by education

<table>
<thead>
<tr>
<th>Income quintiles (stratified by household income per person)</th>
<th>Number of observations</th>
<th>Percentage of people with piped water</th>
<th>Households with piped water stratified by highest education of female members</th>
<th>Full sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom 20th percentile</td>
<td>6581</td>
<td>27.18</td>
<td>768</td>
<td>655</td>
</tr>
<tr>
<td>20–40th percentile</td>
<td>6508</td>
<td>25.40</td>
<td>674</td>
<td>590</td>
</tr>
<tr>
<td>40–60th percentile</td>
<td>6543</td>
<td>26.96</td>
<td>667</td>
<td>560</td>
</tr>
<tr>
<td>60–80th percentile</td>
<td>6694</td>
<td>29.62</td>
<td>660</td>
<td>602</td>
</tr>
<tr>
<td>Top 20th percentile</td>
<td>6904</td>
<td>33.63</td>
<td>665</td>
<td>593</td>
</tr>
<tr>
<td>Full sample</td>
<td>33230</td>
<td>28.62</td>
<td>3434</td>
<td>3000</td>
</tr>
</tbody>
</table>
## Propensity score

**Table 2**  
Logit regression for piped water

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>t-statistic</th>
</tr>
</thead>
</table>

**Village variables**

- Village size (log): 0.08212, 4.269
- Proportion of gross cropped area which is irrigated: > 0.75: -0.04824, -1.185
- Proportion of gross cropped area which is irrigated: 0.5–0.75: 0.19399, 4.178
- Whether village has a day care center: -0.07249, -2.225
- Whether village has a primary school: -0.08136, -1.434
- Whether village has a middle school: -0.09019, -2.578
- Whether village has a high school: 0.26460, 7.405
- Female to male students in the village: 0.10637, 3.010
- Female to male students for minority groups: -0.07661, -2.111
- Main approachable road to village: pucca road: 0.19441, 3.637
- jeepable/kuchha road: -0.00163, -0.033
- Whether bus-stoop is within the village: 0.11423, 2.951
- Whether railway station is within the village: 0.00920, 0.179
- Whether there is a post-office within the village: 0.02193, 0.550
- Whether the village has a telephone facility: 0.33059, 9.655
- Whether there is a community TV center in the village: 0.09859, 2.661
- Whether there is a library in the village: -0.04153, -1.116
- Whether there is a bank in the village: 0.19084, 4.655
- Whether there is a market in the village: 0.31690, 6.092
- Student teacher ratio in the village: 0.00242, 5.295
### Propensity score

**Household variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Propensity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whether household belongs to the Scheduled Tribe</td>
<td>-0.21288</td>
</tr>
<tr>
<td>Whether household belongs to the Scheduled Caste</td>
<td>-0.01045</td>
</tr>
<tr>
<td>Whether it is a Hindu household</td>
<td>-0.24195</td>
</tr>
<tr>
<td>Whether it is a Muslim household</td>
<td>-0.21631</td>
</tr>
<tr>
<td>Whether it is a Christian household</td>
<td>0.40367</td>
</tr>
<tr>
<td>Whether it is a Sikh household</td>
<td>-0.86645</td>
</tr>
<tr>
<td>Household size</td>
<td>0.00337</td>
</tr>
<tr>
<td>Utilization of landholdings: used for cultivation?</td>
<td>0.17109</td>
</tr>
<tr>
<td>Whether the house belongs to the household</td>
<td>-0.18988</td>
</tr>
<tr>
<td>Whether the household owns other property</td>
<td>0.00181</td>
</tr>
<tr>
<td>Whether the household has a bicycle</td>
<td>-0.26514</td>
</tr>
<tr>
<td>Whether the household has a sewing machine</td>
<td>0.01183</td>
</tr>
<tr>
<td>Whether the household owns a thresher</td>
<td>-0.05790</td>
</tr>
<tr>
<td>Whether the household owns a winnower</td>
<td>0.21842</td>
</tr>
<tr>
<td>Whether the household owns a bullock-cart</td>
<td>-0.25900</td>
</tr>
<tr>
<td>Whether the household owns a radio</td>
<td>0.01036</td>
</tr>
<tr>
<td>Whether the household owns a TV</td>
<td>0.08095</td>
</tr>
<tr>
<td>Whether the household owns a fan</td>
<td>0.01336</td>
</tr>
<tr>
<td>Whether the household owns any livestock</td>
<td>-0.07780</td>
</tr>
<tr>
<td>Nature of house: Kuchha</td>
<td>-0.10004</td>
</tr>
<tr>
<td>Pucca</td>
<td>0.12039</td>
</tr>
<tr>
<td>Condition of house: Good</td>
<td>0.00230</td>
</tr>
<tr>
<td>Livable</td>
<td>0.09268</td>
</tr>
<tr>
<td></td>
<td>0.571</td>
</tr>
<tr>
<td></td>
<td>1.914</td>
</tr>
<tr>
<td></td>
<td>-2.854</td>
</tr>
<tr>
<td></td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>-8.243</td>
</tr>
<tr>
<td></td>
<td>0.252</td>
</tr>
<tr>
<td></td>
<td>-0.577</td>
</tr>
<tr>
<td></td>
<td>1.820</td>
</tr>
<tr>
<td></td>
<td>-5.430</td>
</tr>
<tr>
<td></td>
<td>0.251</td>
</tr>
<tr>
<td></td>
<td>1.335</td>
</tr>
<tr>
<td></td>
<td>0.321</td>
</tr>
<tr>
<td></td>
<td>-2.339</td>
</tr>
<tr>
<td></td>
<td>-2.775</td>
</tr>
<tr>
<td></td>
<td>2.709</td>
</tr>
<tr>
<td></td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>1.756</td>
</tr>
</tbody>
</table>
Overlap

Propensity score for households with piped water

Propensity score for households without piped water
Overlap

Prior to matching
- with piped water - average p-score: 0.5495
- without piped water - average p-score: 0.1933

After matching
- with piped water - average p-score: 0.3743
- without piped water - average p-score: 0.3742

650 treatment households lost due to inability to find match
The table below shows the impacts of piped water on diarrhea prevalence and duration for children under five. The data are stratified by household income per capita (quintiles).

### Table 3
Impacts of piped water on diarrhea prevalence and duration for children under five

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Prevalence of diarrhea</th>
<th>Duration of illness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean for those with piped water (st. dev.)</td>
<td>Impact of piped water (st. error)</td>
</tr>
<tr>
<td><strong>Full sample</strong></td>
<td>0.0108 (0.046)</td>
<td>-0.0023* (0.001)</td>
</tr>
<tr>
<td><strong>Stratified by household income per capita (quintiles)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (poorest)</td>
<td>0.0155 (0.055)</td>
<td>0.0032* (0.001)</td>
</tr>
<tr>
<td>2</td>
<td>0.0136 (0.051)</td>
<td>0.0007 (0.001)</td>
</tr>
<tr>
<td>3</td>
<td>0.0083 (0.038)</td>
<td>-0.0039* (0.001)</td>
</tr>
<tr>
<td>4</td>
<td>0.0100 (0.044)</td>
<td>-0.0036* (0.001)</td>
</tr>
<tr>
<td>5</td>
<td>0.0076 (0.042)</td>
<td>-0.0068* (0.001)</td>
</tr>
</tbody>
</table>
Table 3
Impacts of piped water on diarrhea prevalence and duration for children under five

<table>
<thead>
<tr>
<th></th>
<th>Prevalence of diarrhea</th>
<th>Duration of illness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean for those</td>
<td>Impact of</td>
</tr>
<tr>
<td></td>
<td>with piped water</td>
<td>piped water</td>
</tr>
<tr>
<td>(st. dev.)</td>
<td>(st. error)</td>
<td>(st. dev.)</td>
</tr>
<tr>
<td>Illiterate</td>
<td>0.0131</td>
<td>−0.0000</td>
</tr>
<tr>
<td>(0.053)</td>
<td>(0.001)</td>
<td>(1.710)</td>
</tr>
<tr>
<td>At most primary</td>
<td>0.0112</td>
<td>−0.0015</td>
</tr>
<tr>
<td>school educated</td>
<td>(0.045)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>At most</td>
<td>0.0074</td>
<td>−0.0065*</td>
</tr>
<tr>
<td>matriculation</td>
<td>(0.038)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>educated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher secondary</td>
<td>0.0050</td>
<td>−0.0080*</td>
</tr>
<tr>
<td>or more</td>
<td>(0.027)</td>
<td>(0.002)</td>
</tr>
</tbody>
</table>

*Indicates significance at the 5% level or lower.
Conclusions

- They find significantly lower prevalence and duration of the disease for children living in households with piped water as compared to a comparison group of households matched on the basis of their propensity scores.
- They find no evidence of significant gains for the poorest 40% in terms of incomes.
- Health gains from piped water tend to be lower for children with less well-educated women in the household.
- Income poverty and lack of education and knowledge may well constrain the potential health gains from water infrastructure improvements.
- The incidence of health gains need not favor children from poor families even when facility placement is pro-poor.