Evaluation of Development Programs

Randomized Controlled Trials or Regressions?

Chris Elbers
Jan Willem Gunning

The World Bank
Development Economics Vice Presidency
Partnerships, Capacity Building Unit
September 2013
Abstract

Can project evaluation methods be used to evaluate programs: complex interventions involving multiple activities? A program evaluation cannot be based simply on separate evaluations of its components if interactions between the activities are important. In this paper a measure is proposed, the total program effect (TPE), which is an extension of the average treatment effect on the treated (ATET). It explicitly takes into account that in the real world (with heterogeneous treatment effects) individual treatment effects and program assignment are often correlated. The TPE can also deal with the common situation in which such a correlation is the result of decisions on (intended) program participation not being taken centrally. In this context RCTs are less suitable even for the simplest interventions.

The TPE can be estimated by applying regression techniques to observational data from a representative sample from the targeted population. The approach is illustrated with an evaluation of a health insurance program in Vietnam.
Evaluation of Development Programs:

Randomized Controlled Trials or Regressions?

Chris Elbers and Jan Willem Gunning*

JEL Classification Codes: C21, C33, O22

Keywords: program evaluation; randomized controlled trials; policy evaluation; treatment heterogeneity; budget support; sector-wide programs; aid effectiveness

Sector Board: Economic Policy (EPOL)

* The authors are professors at the VU University Amsterdam and fellows of the Tinbergen Institute. Their addresses are c.t.m.elbers@vu.nl and j.w.gunning@vu.nl (corresponding author). They are grateful to Remco Oostendorp, Menno Pradhan, Martin Ravallion, Elisabeth Sadoulet, Finn Tarp, the editors and two anonymous referees of the World Bank Economic Review and to seminar participants in Amsterdam, Namur, Oxford and Paris for very valuable comments on previous versions.
Experimental methods for impact evaluation presuppose that the intervention is well-defined: the “project” is limited in space and scope (e.g. Duflo et al., 2008). However, governments, NGOs and donor agencies are often interested in evaluating the effect of a program consisting of various interventions, e.g. a sector-wide health or education programs (De Kemp et al., 2011). Program evaluation faces two complications. First, a sharp distinction between treatment and control groups is usually impossible. For example, a program in the education sector may involve activities such as school building, teacher training and supply of textbooks. Typically all communities are affected in some way by the program, but they may differ dramatically in what interventions they are exposed to and the extent of that exposure. Secondly, in a program the interventions are typically implemented at various administrative levels so that the policy maker has only imperfect control over actual treatment.

The impact of such a program cannot simply be calculated on the basis of the results of randomized controlled trials (RCTs). This would run into well known problems of external validity (Bracht and Glass, 1968, Rodrik, 2008, Ravallion, 2009, Banerjee and Duflo, 2009, Deaton, 2010, Imbens, 2010) even if the program involved only a single intervention. In addition, if the program involves multiple interventions and interactions are important then it is not clear how RCT evaluations of individual components of the program should be combined to an overall assessment of the program. However, regression techniques can be used for program evaluation. This involves drawing a representative sample of beneficiaries (e.g. households, schools, communities) and collecting data on the combination of interventions experienced by each beneficiary, together with other possible determinants of the outcome variables of interest. Regression techniques can then be used to estimate the impact of the various interventions.¹ In this paper this approach is generalized by allowing for treatment heterogeneity and a way of estimating aggregate program impact is proposed.
Obviously, the intervention variables are likely to be endogenous in a regression analysis. For example, an unobserved variable such as the political preferences of the community may affect both the impact variable of interest and the intervention. Also, the impact of the intervention will differ between beneficiaries and the allocation of interventions across beneficiaries may be based on such treatment heterogeneity, either through self-selection or through the allocation decisions of program officers. Heckman (1997) and Heckman et al. (2008) call this “selection on the gain”. The first complication is usually dealt with by using panel data or by randomized assignment of treatment. The second complication is much more serious. It may be particularly hard for RCTs when program assignment in practice cannot be mimicked by assignment to the treatment arm in an RCT since this would not capture the way program officers take their decisions. However, it will be shown that regression techniques can be adapted so as to produce an appropriate estimate of the program effect.

The paper is organized as follows. In the first section the total program effect (TPE) is introduced. This measure extends the average treatment effect on the treated (ATET). The TPE is suitable for complex interventions and can deal with selection on the gain (treatment heterogeneity). Then two complications are considered: correlation between program variables and the controls in section 2 and spillover effects in section 3. Section 4 investigates whether estimating the TPE using RCTs is an alternative. The approach is illustrated in section 5 by estimating the TPE for a health insurance intervention in Vietnam. Section 6 concludes.

I. The Total Program Effect (TPE)

Consider the following model:

\[ y_{it} = \alpha X_{it} + \beta_i P_{it} + \gamma_i + \eta_i + \varepsilon_{it} \]  

(1)
where \( y \) measures an outcome of interest, in this paper taken to be a scalar; \( t = 0, 1 \) is the time of measurement; and \( i = 1, \ldots, n \) denotes the unit of observation, e.g. households or locations. \( P \) denotes a vector of the interventions to be evaluated and \( X \) a vector of observed controls.\(^2\) The \( P \)-variables can either be binary variables or multi-valued (discrete or continuous) variables. \( \alpha \) and \( \beta \) are vectors of parameters, \( \gamma \) denotes a time effect and \( \eta \) represents time-invariant unobserved characteristics and \( \epsilon \) is the error term, assumed to be independent over time. It is also assumed that the interventions and control variables are uncorrelated with the error process:

\[
X_{it}, X_{i0}, P_{it}, P_{i0} \perp \epsilon_{it}, \epsilon_{i0}.
\]

At this stage \( P \) and \( X \) are assumed to be independent:

\[
X_{it}, X_{i0} \perp P_{it}, P_{i0}.
\]

This will be relaxed in section 2. Note that equation (1) excludes spillover effects of the type where \( y_{it} \) depends on \( P_{ji} \) (\( i \neq j \)) and \( j \) is not necessarily included in the sample. This point will be discussed in section 3. In many applications (1) will represent a reduced form or “black box” regression, but it can also represent a structural model.

The evaluator is interested in the expectation (in the population) of the effect of interventions on the outcome variable, the total program effect (TPE):\(^3\)

\[
\text{TPE} = E \beta_i (P_{it} - P_{i0}).
\]

Note that the impact parameters \( \beta_i \) need not be the same for all \( i \): heterogeneity of program impact is allowed.

As an example consider a very simple special case:

\[
y_{it} = \beta_i P_{it} + \gamma_i + \eta_i + \epsilon_{it} \quad t = 0, 1
\]
where \( P_i \) now is a binary variable rather than a vector, \( P_{i0} = 0 \) for all \( i \) and \( P_i = P_{i1} - P_{i0} \). Taking first differences gives:

\[
\Delta y_i = \beta P_i + \gamma + \Delta \varepsilon_i
\]

where \( \gamma = \gamma_1 - \gamma_0 \). This is analogous to the equation for a standard project evaluation, but written in differences.\(^4\) The TPE for this case equals \( E\beta_i P_i \) which is related to the familiar average treatment effect on the treated (ATET)

\[
\text{ATET} = \frac{\text{TPE}}{EP_i}.
\]

In another special case of equation (1) the TPE can be identified as follows. Assume that data are available from a random sample and that for a subsample (the “control group”) there is no change in the interventions: \( P_{i1} = P_{i0} \). (At this stage it is not assumed that the assignment to intended “treatment” and “control” groups is random.) Taking first differences in (1) for this group gives:

\[
\Delta y_i = \alpha \Delta X_i + \gamma + \Delta \varepsilon_i \quad \text{if} \; P_i = 0.
\]

This allows estimation of \( \alpha \) and hence \( \hat{\alpha} \Delta X_i \) so that the TPE can be estimated as

\[
\hat{\text{TPE}} = \overline{\Delta y_i} - \hat{\alpha} \overline{\Delta X_i}.
\]

However, in a program consisting of multiple interventions, the context of this paper, there will usually not be a sufficiently large control group to make this identification strategy realistic. Indeed, typically the control group will be empty: all \( i \) will have experienced a change in at least some components of the vector \( \Delta P_i \).

For this more general case

\[
\Delta y_i = \alpha \Delta X_i + \beta \Delta P_i + \gamma + \Delta \varepsilon_i
\]  

(3)
Allowing for “selection on the gain”, correlation between impact parameters $\beta_i$ and the program variables $P_i$ and also for correlation between $\beta_i$ and $X_i$ equation (3) can be rewritten as

$$\Delta y_i = \alpha \Delta X_i + E(\beta_i \mid \Delta X_i, \Delta P_i) \Delta P_i + \gamma + \omega_i,$$  \hspace{1cm} (4) \hspace{1cm}

where $\omega_i = \Delta e_i + (\beta_i - E(\beta_i \mid \Delta X_i, \Delta P_i)) \Delta P_i$ and this is uncorrelated with $\Delta X_i$ and $\Delta P_i$.

The term $E(\beta_i \mid \Delta X_i, \Delta P_i)$ can be approximated linearly: \hspace{1cm} $E(\beta_i \mid \Delta X_i, \Delta P_i) \approx \delta_0 + \delta_1 \Delta X_i + \delta_2 \Delta P_i$.

Substitution in (4) and collecting terms gives

$$\Delta y_i = \gamma + \theta_0 \Delta X_i + \theta_1 \Delta P_i + \theta_2 \Delta X_i \otimes \Delta P_i + \theta_3 \Delta P_i \otimes \Delta P_i + \omega_i$$ \hspace{1cm} (5) \hspace{1cm}

where

$$\theta_0 \Delta P_i + \theta_1 \Delta X_i \otimes \Delta P_i + \theta_2 \Delta P_i \otimes \Delta P_i$$

is the approximation of $T_i = E(\beta_i \Delta P_i \mid \Delta X_i, \Delta P_i)$.

Equation (5) can be estimated using the sample data. The estimated coefficients can then be used to estimate $T_i$ as

$$\hat{T}_i = \hat{\theta}_0 \Delta P_i + \hat{\theta}_1 \Delta X_i \otimes \Delta P_i + \hat{\theta}_2 \Delta P_i \otimes \Delta P_i.$$

The TPE can now be estimated as the average of $\hat{T}_i$ in the sample.

$$\hat{TPE} = \frac{1}{n} \sum_i \hat{T}_i = \hat{\theta}_0 \overline{\Delta P_i} + \hat{\theta}_1 \overline{\Delta X_i \otimes \Delta P_i} + \hat{\theta}_2 \overline{\Delta P_i \otimes \Delta P_i}$$ \hspace{1cm} (6) \hspace{1cm}

where bars denote sample averages.
In practice this means that one regresses $\Delta y_i$ on $\Delta X_i$, $\Delta P_i$ and their interactions with $\Delta P_i$ and collects all terms involving $\Delta P_i$ to calculate the total program effect. Since the estimated TPE is linear in the $\hat{\theta}$ parameters its standard error can be obtained from the covariance matrix of the OLS-coefficients.

It is instructive to consider the special case of equation (5) where $D_i = \Delta P_i$ is a binary variable taking the value 1 for the treatment group and 0 for the control group, i.e. the case of a difference-in-difference analysis. Equation (5) now reduces to

$$\Delta y_i = \gamma + \theta_1 \Delta X_i + \theta_2 D_i + \theta_3 D_i \Delta X_i + \omega_i$$

since in this case $D_i^2 = D_i$. Compared to a standard diff-in-diff regression this equation contains the interaction term $D_i \Delta X_i$.

The total program effect will in this case be estimated as

$$\hat{\text{TPE}} = \hat{\theta}_2 \bar{D}_i + \hat{\theta}_3 \bar{D}_i \bar{\Delta X}_i.$$  \quad (7)

This shows that when the sample is representative sample means can be used to construct the total program effect. The interaction term in (7) avoids the bias resulting from correlations between treatment effects and either program participation or controls.

Many diff-in-diff studies do not include the interaction terms (e.g., Khandker et al., 2009 or Almeida and Galasso, 2010). Studies that do often report estimates of impact for different values of the controls $X$ which makes it difficult to assess the aggregate impact of a program.
Equation (1) allows for two types of selection effects: $P_\alpha$ may be correlated with $\beta_\iota$ or with the unobserved characteristics $\eta_\iota$. A correlation of $P_\alpha$ and $\eta_\iota$ is dealt with by differencing, as in (3). However, the TPE measures the effect of the program inclusive of selectivity in the assignment of program interventions resulting in a correlation of $\beta_\iota$ and $\Delta P_\iota$. This is appropriate since the way the program was assigned (in an ex post evaluation) or will be assigned (in an ex ante evaluation) is one of its characteristics. If the program was successful in part because program officers made sure the program interventions were assigned to households or locations where they expected a high impact, then obviously the evaluation should reflect this. In fact the evaluation would be misleading if it tried to “correct” for such selection effects by presenting (if this were feasible) an estimate ($E\beta_\iota$) of the program’s impact if it had been assigned randomly.

Recall that in the special, binary case of a ‘project’ evaluation $TPE = E\beta_\iota \Delta P_\iota = ATET \times E\Delta P_\iota$. If administrative data can be used to estimate $E\Delta P_\iota$ the question arises whether the ATET is identified in an RCT. Obviously this is the case if $\beta_\iota = \beta$ for all $\iota$. More generally, if $\Delta P_\iota$ and $\beta_\iota$ are independent the TPE can be estimated on the basis of an RCT: the trial would give an estimate of $E\beta_\iota$ which in this case is also the ATET. A special case of independence is that of universal treatment ($P_\iota = 1$ for all $\iota$). In the most general case when $\Delta P_\iota$ and $\beta_\iota$ are not independent the ATET as established by an RCT may differ from the ATET in the population and estimating the TPE on the basis of RCTs can become problematic. This issue will be considered in section 4.
II. Correlation between $P$ and $X$

In the previous section $P$ and $X$ were assumed to be independent. $(P, X)$ correlations are often important in evaluations. For example, changes in teacher training may induce changes in parental input. Not all such inputs will be observed (e.g. additional parental help with homework will probably not be recorded); $P_i$ will then be correlated with $\beta_i$ and this was already considered in the previous section. Conversely, if the parental input is observed then $P_i$ will be correlated with $X_i$. In that case the TPE identifies the direct effect of $P$, but not its total effect (including the indirect effect through induced changes in $X$). If the induced effect is to be included then the affected components of $\Delta X_i$ should be omitted from the regression (5).

If causality is in the reverse direction, from $\Delta X_i$ to $\Delta P_i$, then there is no need to amend the section 1 estimate of the TPE since there is no induced change in $\Delta X_i$. (The asymmetry arises because in either case the interest is in the impact of changes in $\Delta P_i$, rather than in the impact of changes in $\Delta X_i$.)

In the general case where the direction of causality is not known it will usually not be possible to estimate the indirect effect of the program. Occasionally, however, appropriate instruments can be found so that the impact of $\Delta P_i$ on $\Delta X_i$ can be identified.
III. Spillover Effects

Recall that in section 1 spillover effects were excluded: in equation (1) $y_i$ of case $i$ does not depend on $P_j$ of case $j$. In evaluations there are two important situations where this assumption is untenable. First, Chen et al. (2009) and Deaton (2010) discuss the possibility that policy in control villages is partly determined by policies in treatment villages so that the SUTVA (stable unit treatment value assumption) is violated. Indeed, if policies thus affected are not represented in the policy vector $P_i$ this creates a classical case of omitted variable bias. In Chen et al. the problem arises because the data record participation in a particular program as a binary $P_i$ variable, while other programs which may affect the outcome are initially ignored. In the approach advocated in the present paper all potentially relevant programs would in principle be included in $P_i$ so that the problem of SUTVA violation is avoided. Secondly, policies in village $j$ may affect outcomes in village $i$. For example, a program aimed at an infectious disease in village $j$ may affect health outcomes in the “untreated” village $i$. If the external effects of policy are general equilibrium effects such as regional wage increases, it will be hard to identify the full impact of a policy. But often more structure can be imposed, e.g. by including a proxy for relevant policies in neighboring villages in the outcome regression, so that equation (3) is extended to

$$\Delta y_i = \beta_i \Delta P_i + \alpha \Delta X_i + \gamma + \delta \Delta K_i + \Delta \varepsilon_i,$$

where $\Delta K_i$ is the proxy for policy changes in the neighborhood. If there is sufficient variation in $K_i$ then $\delta$ is identified in this regression. The TPE would then be $E \beta_i \Delta P_i + \delta E \Delta K_i$. 


IV. Regression Methods and RCTs Compared

In section 1 it was shown how the TPE can be estimated using regression methods. A natural question is whether the TPE can also be estimated using RCTs. Using RCTs may be difficult, e.g. because in programs the distinction between treatment and control groups may break down. However, there may be problems even in the case of binary treatments, namely under treatment heterogeneity when the probability of treatment is correlated with the individual impact parameters $\beta_i$ and unknown to the evaluator. If this correlation arises through self-selection then the usual response is to consider the average treatment effect on the treated rather than the average treatment effect in the population. If, however, the correlation arises at a higher level, e.g. because the policy maker targets on observables, then an RCT would have to mimic this assignment, possibly by stratifying the sample on the basis of the targeting variables.

But in many government and NGO programs the “policy maker” does not directly control the $P$ variables: assignment is decided by lower level staff (“program officers”) on the basis of private information, variables that cannot be observed by the policy maker or the evaluator. In this case an RCT can still identify the TPE, but at the cost of having to randomize at a higher level than the treatment under consideration: randomization would apply to program officers rather than beneficiaries. This implies that the power of the statistical analysis may be reduced. It also involves losing the direct link with the intervention.

This may be illustrated with an example. Consider the following model

$$y_i = \beta_i P_i + \gamma + \epsilon_i$$

where $\beta_i$ and $\epsilon_i$ are independent, $P_i$ is binary and $E\epsilon_i = 0$. For simplicity $\beta_i$ will be considered as the intention-to-treat impact, so that a subject $i$’s refusal to undergo offered treatment $P_i$ is
reflected in $\beta_i$, rather than in $P_i$. Program implementation involves program officers who have imperfect knowledge of $\beta_i$: they perceive $\omega_i = \beta_i + \eta_i$ and will assign treatment if and only if $\omega_i > 0$. Assume that $\eta_i$ has mean zero and is independent of $\beta_i$ and $\epsilon_i$. Crucially, this knowledge of program officers is unknown to the evaluator. Denote the CDF of $\eta_i$ by $F$. With this assignment rule $P_i$ is exogenous (i.e. independent of $\epsilon_i$).

An RCT evaluation might involve drawing a random sample from the population and assign treatment randomly within this sample. The researcher would then estimate the program’s intention to treat effect (ITE) as $E\beta_i$. The TPE would be estimated as $E\beta_i E P_i$.

This would be incorrect since, under the assumptions made above

$$TPE = E\beta_i P_i = E(\beta_i | \beta_i + \eta_i > 0)P(\beta_i + \eta_i > 0) = E[(1 - F(-\beta_i))\beta_i] \neq EP_i E\beta_i.$$ 

(Note that $E(1 - F(-\beta_i)) = P(\beta_i + \eta_i > 0) = EP_i$. As before, the ATET = TPE / $EP_i$.) The problem arises because in this case the RCT design does not mimic the actual assignment process. To obtain an unbiased estimate of the TPE randomization would have to take place at a higher level, that of the program officers. The control group then consist of program officers who never “treat” and the treatment group of program officers who sometimes (but not always) treat.

The proposed regression method gives an unbiased estimator of the TPE using observational data for $(y_i, P_i)$ from a random sample of the population. The difference is that while the RCT approach compares average outcomes at the level of program officers the regression approach does so at the level of beneficiaries. The RCT approach therefore has lower statistical power.
Moving beyond the example there is a more fundamental objection to the RCT approach if outcomes depend not only on $P$ but also on $X$, as in (1). If the RCT involved randomization over actual program officers then it is unlikely that randomization can also be achieved in terms of all the confounding $X$ variables since program officers will not have been posted randomly across space. This introduces a correlation between $X$ and characteristics of the program officers and hence a correlation between $P$ and $X$. The two groups of program officers (“treatment” and “control”) will therefore differ systematically so that internal validity is lost. The proposed approach, by contrast, collects data at the level of beneficiaries and can therefore control for differences in $X$.

In summary, estimating the TPE on the basis of group averages from RCTs becomes problematic when $\beta$ and $P$ are correlated as a result of targeting on the basis of unobservables. If one randomizes at the level of beneficiaries the TPE estimator will be biased because the correlation is not taken into account. If one randomizes at the level of program officers the estimator is inefficient and, if confounders are important, may become inconsistent.

V. An Empirical Example: Estimating the Total Program Effect for a Health Insurance Program in Vietnam

To illustrate how the total program effect can deviate from a naïve approach to calculating the effect of a program a study of the impact of a health insurance program in Vietnam (Wagstaff and Pradhan, 2005) is reconsidered. Health insurance was introduced between the 1992-93 and the 1997-98 rounds of the Vietnam Household Living Standards Survey (General Statistics Office of Vietnam, 1993 and 1998). To account for possible treatment heterogeneity Wagstaff and Pradhan match households on propensity scores and then compare changes in health
outcomes (as well as some non-health outcomes) between insured households or individuals and (matching) uninsured households or individuals. They find modest favorable effects on children’s nutritional status, a mild effect on health expenditure and a sizeable effect on non-health spending.

A propensity score based approach is not suitable for calculation of a total program effect since the common support requirement in a PSM approach will exclude part of the population in a systematic way. Therefore the Vietnam data are used to estimate the effect of the program using a standard diff-in-diff approach i.e. without allowing for heterogeneity (labeled ‘naïve’). The results are compared with an estimate of the TPE. In this case the ‘program’ is a simple intervention. This makes a comparison with a standard approach clearer.

The data are summarized in Table 1. A difficulty is that some of the outcome variables are individual anthropometric measurements while only households can be matched between survey rounds. Therefore the individual measurements have been averaged per household - a crude procedure only suitable for the current purpose of illustrating the TPE. Lacking information on 1992-3 the sampling weights from 1997-8 are used; clustering is also based on the 1997-8 survey round.

The outcome variables considered are changes in arm circumference, height, body weight, health expenditure and total expenditure. The explanatory variables are the other variables shown in Table 1 (insurance status and the controls school attended, currently attending school, gender, age, a farm dummy, household size): and their interaction with the intervention variable are used as explanatory variables. When total expenditure is not a dependent variable it is also used as control variable.
Table 2 summarizes the results. First a naïve regression is run (without interaction terms) and the implied program effect (calculated as the regression coefficient of insurance times mean insurance). This naïve program effect is then compared with a TPE calculated as in equation (6).

The results show striking differences between the two methods. In the case of arm circumference the standard method would have led to the conclusion that insurance had no (significant) effect. Once treatment heterogeneity is allowed for the effect is in fact highly significant albeit very small. For height neither method finds a significant effect. For body weight both methods show a significant increase but the effect is more than twice as large when heterogeneity is allowed for.

Insurance appears to have no significant effect on health expenditure irrespective of the method used. Both methods do find a substantial (and significant) effect of insurance on total consumption. Again, the effect is stronger once one takes heterogeneity into account.

Obviously, there is no reason why these results should generalize. However, they do suggest that treatment heterogeneity can have a substantial effect on the estimates of a program’s impact. A simple way to investigate this possibility is to test for the joint significance of the coefficients on the variables which would not normally be included in the regression: the interactions of treatment variables with themselves and with the controls. When this test indicates that heterogeneity may be an issue it is advisable to calculate the TPE.

**VI. Conclusion**

Policy makers in developing countries, NGOs and donor agencies are under increasing pressure to demonstrate the effectiveness of their program activities. At the same time there is a growing
interest in using randomized controlled trials (RCTs) for impact evaluation of projects. This raises the question to what extent RCTs can be used to evaluate programs, for instance by aggregating the impact of the components of the program. This question is particularly relevant for the evaluation of budget support or of NGOs which typically involve a wide variety of activities.

The strength of RCTs is in establishing proof of principle. Going further and using RCTs to estimate the impact of programs is possible in special cases but becomes problematic if the probability of assignment is correlated with the effectiveness of the intervention. For example, teachers may give more attention to children who they think can benefit more from it. An RCT which randomizes at the level of beneficiaries (children) would produce a biased estimate of the program effect by ignoring this correlation between assignment and treatment effects. Alternatively, randomization at the appropriate level (teachers) would require a larger sample for the same precision. If confounders are important and correlated with characteristics of the program officers, the RCT-based estimate of the program’s impact would even be inconsistent.

The approach proposed in this paper requires observational panel data for a representative sample of beneficiaries rather than experimental data for randomly selected treatment and control groups. If treatment is exogenous this will correctly reflect the assignment process even under treatment heterogeneity. Instead of estimating average impact coefficients for each of the various interventions of the program, the expected value (across beneficiaries) of the total impact of the combined interventions is estimated. This gives the total program effect (TPE). The paper has shown how and under what conditions regression techniques can be used to estimate the TPE in the presence of selection effects. As an example TPE estimates for a simple intervention: a health insurance program in Vietnam were presented. The example shows that
allowing for heterogeneity can lead to very different estimates of a program’s effect. The proposed method offers a simple way of dealing with such heterogeneity.

The approach has three advantages. First, by using observational data for a random sample from the population of intended beneficiaries external validity is ensured. While the disadvantages of observational data are well known, this is an important advantage. Secondly, by focusing on the combined effect of program components they are automatically correctly weighted. Finally, it avoids the problems which RCTs encounter when assignment is imperfectly controlled and correlated with unobservables, as is plausible in development programs.
Notes

1. This approach is discussed in White (2006) and Elbers et al. (2009).

2. Here $P$ reflects “actual” treatment. In principle it could reflect “intended” treatment if intended treatment can be observed, e.g. because intended beneficiaries were offered vouchers.

3. Strictly speaking this is the total effect of changes in the program. The symbol $E$ is used for population averages and a bar over a variable for sample averages. Note that the total program effect does not include general equilibrium effects of the program.

4. This assumes that the autonomous trend $\gamma = \gamma_1 - \gamma_0$ is the same for all subjects (or, alternatively that the difference $\Delta \gamma$ is exogenous and can be treated as part of the residual). In the terminology of double differencing this is the assumption of parallel trends. If this assumption is questionable then data for more periods are needed to estimate how trends depend on $P$. This paper abstracts from this complication and limit the analysis to two periods. The extension to more periods is non-trivial but conceptually straightforward.

5. Higher order approximations would not change the argument but it should be noted that the number of regressors expands very rapidly. De Janvry et al. (2012) account for treatment heterogeneity in a similar way in the context of a schooling program.

6. Obviously, to identify $\theta$ a restriction on parameters like $\theta_{4/k} = \theta_{4/k}$ is required.

7. Differencing is sufficient because of the assumption of parallel trends (cf. footnote 5).

8. Imbens (2010) describes a reduction in class size in all California schools. This is an example of universal treatment.

9. Deaton (2010) gives the example where random assignments made by the central government (e.g. the Ministry of Education) are partly offset by induced changes in allocations by local or provincial governments. Ravallion (2012) gives a similar example and Chen et al. (2009) quantify such a spillover effect in China. Similarly, the political economy may be such that the central government is unable to prevent allocations being diverted to favored ethnic or political groups. In either case $P_i$ might be correlated with $\beta_i$.

10. This is similar to the case considered by Das et al. (2004, 2007) where teacher absenteeism as a result of HIV/AIDS induces greater parental input.

11. Recall that the approach does not involve a distinction between treatment and control groups: most if not all subjects receive some treatment.
This has implications for sampling: since data on policies in neighboring villages are required one must sample groups (possibly pairs) of adjacent villages.

Duflo et al. (2008, pp. 3935-37) make this point in a similar context (partial compliance) concluding that “One must compare all those initially allocated to the treatment group to all those initially randomized to the comparison group”.

This is shown in the supplemental appendix.

It should be noted that the intervention variable is not binary (as it would be in a ‘project’) since insurance enrollment is measured as an average at the household level.
References


Table 1: Data for the Vietnam Insurance Example

<table>
<thead>
<tr>
<th>Variable: change in (average)</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arm circumference (cm)</td>
<td>1.154</td>
<td>2.013</td>
<td>-7.3</td>
<td>9.4</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>5.175</td>
<td>11.35</td>
<td>-49.57</td>
<td>39.84</td>
</tr>
<tr>
<td>Body weight (kg)</td>
<td>2.983</td>
<td>6.544</td>
<td>-27.75</td>
<td>26.25</td>
</tr>
<tr>
<td>Health expenditure ('000 Dong)</td>
<td>1,081</td>
<td>5,519</td>
<td>-8808</td>
<td>23,3965</td>
</tr>
<tr>
<td>Total consumption expenditure ('000 Dong)</td>
<td>6,513</td>
<td>8,009</td>
<td>-22,988</td>
<td>11,6826</td>
</tr>
<tr>
<td>Insurance (binary at individual level)</td>
<td>0.170</td>
<td>0.268</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>School attended(^{16})</td>
<td>-0.017</td>
<td>0.683</td>
<td>-3.5</td>
<td>3</td>
</tr>
<tr>
<td>Currently attending school (binary at individual level)</td>
<td>0.082</td>
<td>0.388</td>
<td>-2</td>
<td>2</td>
</tr>
<tr>
<td>Gender</td>
<td>0.002</td>
<td>0.138</td>
<td>-0.75</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td>3.522</td>
<td>8.299</td>
<td>-48.43</td>
<td>48.6</td>
</tr>
<tr>
<td>Farm dummy</td>
<td>-0.079</td>
<td>0.421</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Household size</td>
<td>-0.267</td>
<td>1.696</td>
<td>-18</td>
<td>11</td>
</tr>
</tbody>
</table>

The number of observations varies between 4299 and 4305.

Table 2: Total Program Effects

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Naïve program effect† (I) (s.e.)</th>
<th>Total program effect†† (II) (s.e.)</th>
<th>R-squared of underlying regressions I</th>
<th>II</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arm circumference</td>
<td>.022 (.029)</td>
<td>0.090*** (0.027)</td>
<td>0.22</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>-0.190 (0.154)</td>
<td>.095 (0.139)</td>
<td>0.34</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Body weight</td>
<td>0.167* (0.083)</td>
<td>0.384*** (0.074)</td>
<td>0.31</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>Health expenditure</td>
<td>-28.08 (60.59)</td>
<td>-52.79 (51.01)</td>
<td>0.03</td>
<td>0.04</td>
<td>Total consumption included in controls</td>
</tr>
<tr>
<td>Health expenditure</td>
<td>55.41 (66.42)</td>
<td>64.32 (52.87)</td>
<td>0.00</td>
<td>0.00</td>
<td>Total consumption expenditure not included</td>
</tr>
<tr>
<td>Total consumption expenditure</td>
<td>626.7*** (110.9)</td>
<td>888.8*** (105.7)</td>
<td>0.10</td>
<td>0.12</td>
<td>Total consumption expenditure not included</td>
</tr>
</tbody>
</table>

Robust clustered standard errors in parentheses. In all but the health expenditure regressions squared intervention and interactions of controls with intervention are jointly significant. Significance: * indicates 5% threshold, *** 0.1%.
† The naïve program effect is calculated as the regression coefficient on the insurance variable time the estimated population mean of that variable.
†† The total program effect is calculated according to equation (6).
The sampling errors on the estimated population means are not taken into account.
Using RCTs

“Program Officers” (POs) are divided into treatment- and control-POs. All subjects within the catchment area of a treatment-PO are considered as treated (i.e., we want to estimate the intention to treat effect).

Consider the following model linking outcome \( y_{ij} \) to (actual) treatment \( P_{ij} \):

\[
y_{ij} = \alpha_i + \beta_i P_{ij} + \epsilon_{ij},
\]

where \( i \) refers to the program officer responsible for administrating treatment to subject \( j \) who falls within the catchment area of \( i \). The disturbance \( \epsilon_{ij} \) is assumed to be homoscedastic and independent of \( \alpha_i, \beta_j \) and \( P_{ij} \). To model clustering by POs an officer random effect \( \alpha_i \) is included in the model. Random effects are assumed to be i.i.d. and independent of \( \beta_j \) and \( P_{ij} \).

We further assume that the number of subjects per PO is constant to avoid trivial complications of weighing.

The evaluator wants to estimate \( \operatorname{TPE} = E \beta_j P_j \) and in order to capture any selectivity in application of treatment by the program officers (PO) a random sample of POs has been drawn and subsequently been randomly divided into a group \( T \) of treatment-POs who are supposed to apply treatment to the ultimate beneficiaries \( j \) and a group \( C \) of control-POs who are asked not

\(^1\) The context is that of section 4 in the main text of the paper.
to give treatment to subjects. Within the catchment area of sampled POs a random sample of subjects is drawn for whom we observe (at least) \( y_{ij} \). This allows estimation of the TPE as the difference in average outcomes between group \( T \) and group \( C \) subjects: 

\[
\hat{TPE} = \bar{y}_T - \bar{y}_C = \bar{\alpha}_T - \bar{\alpha}_C + \left[ \beta \bar{y}_P \right]_T + \bar{\epsilon}_T - \bar{\epsilon}_C, \tag{A.1}
\]

where the bars denote sample averages over the two groups of subjects. Since this estimator is unbiased, its precision can be determined by the variance:

\[
\text{MSE}(\hat{TPE}) = \left( \frac{1}{n_T} + \frac{1}{n_C} \right) \sigma^2 + \frac{1}{N_T} \left[ \text{var}(\beta y_P) \right]_T + \left( \frac{1}{N_T} + \frac{1}{N_C} \right) \sigma^2 \epsilon
\]

where \( n_T \) and \( n_C \) denote the number of sampled treatment-POs and control-POs, \( N_T \) the total number of sampled subjects associated with treatment-POs, and \( N_C \) the number of sampled subjects falling under control-POs.

**Regression using observational data**

Now consider sampling directly at the level of subjects. Typically such a sample will also be clustered, albeit not necessarily by PO. To create a ‘level playing field’ we will assume that the sample has \( n = n_T + n_C \) clusters with a total of \( N = N_T + N_C \) subjects. For each sampled subject \( j \) from cluster \( i \) we observe \( P_y \) (actual treatment) and \( y_{ij} \). The estimator for the TPE reduces to

\[
\hat{TPE} = \bar{y} - \frac{y_i (1 - P_y)}{1 - P_y} = \beta \bar{y}_P + \alpha_i (1 - P_y) + \epsilon - \frac{\epsilon_i (1 - P_y)}{1 - P_y}. \tag{A.2}
\]

Assuming as in the RCT setup that \( \alpha \) is independent of \( P_y \) and \( \beta \) this estimator is again unbiased\(^2\) and

\(^2\)Correlation of \( \alpha \) and \( \beta \) would reflect level effects which, as explained in section 2, should be neutralized by using differenced data.
\[
\text{MSE}(\hat{\text{TPE}}) = \frac{1}{N} \text{var}(\beta_j P_j) + \text{var} \left( \alpha_i = \frac{\alpha_i (1-P_y)}{1-P_y} \right) + \text{var} \left( \varepsilon_y = \frac{\varepsilon_y (1-P_y)}{1-P_y} \right).
\]

Using the delta method and the equality \( E(P_y - \overline{P}_y)^2 = \frac{N-1}{N} EP(1-EP) \) it can be verified that\(^3\)

\[
\text{var} \left( \varepsilon_y = \frac{\varepsilon_y (1-P_y)}{1-P_y} \right) = \text{var} \left( \frac{1}{N} \sum_y \varepsilon_y (P_y - \overline{P}_y) \right) \approx \frac{\overline{P}_y}{1-\overline{P}_y} \frac{1}{N-1} \sigma^2,
\]

and likewise that

\[
\text{var} \left( \alpha_i = \frac{\alpha_i (1-P_y)}{1-P_y} \right) = \text{var} \left( \frac{1}{N} \sum_y \alpha_i (P_y - \overline{P}) \right) \approx \frac{\overline{P}_y}{1-\overline{P}_y} \frac{1}{N-1} \sigma^2.
\]

It follows that in the regression setup precision is of order \( \sqrt{N} \) while in the RCT setup precision if at best of order \( \sqrt{N_T/2} \) and, if clustering of data is an issue, of order \( \sqrt{n_T/2} \). (Note that if the two groups are of equal size: \( N_T = N/2 \), then the regression setup is twice as precise as the RCT setup.)

**Covariates**

Both methods fail if \( \varepsilon \) and \( \beta P \) are correlated. What if there are observables \( X_y \) determining both \( P \) and \( y \)? This could be the result of program targeting. In that case formulas (A.1) and (A.2) can no longer be used. To account for the confounding effect of covariates a regression approach is required, also with an RCT setup. For RCTs using intention to treat by PO for estimating the TPE, efficient estimation would amount to a regression equation like

\[
y_j = \alpha_i + \text{TPE} I_{(i \in R)} + \gamma X_y + \varepsilon_y.
\]

The reason formula (A.1) can no longer be used is that randomization over POs does not guarantee randomization over observables \( x_y \). Applying formula (A.1) we would find\(^3\)

\[3 \text{ In this case } E \text{ denotes an average over all possible samples.}\]
\[ \hat{TP}E = \bar{y}_r - \bar{y}_c = \bar{\alpha}_r - \bar{\alpha}_c + \gamma (\bar{X}_r - \bar{X}_c) + [\beta_0 P]_r T + \bar{\varepsilon}_r - \bar{\varepsilon}_c. \]

The bias \( \gamma (\bar{X}_r - \bar{X}_c) \) would vanish if \( \bar{X}_r \approx \bar{X}_c \), i.e., when \( X_\theta \) and \( I_{\{i=aT\}} \) are uncorrelated.