The Effectiveness of International Aid: A Generalized Propensity Score Analysis

Jing Liu
Department of Agricultural Economics, Purdue University
liu207@purdue.edu

Raymond J.G.M. Florax
Department of Agricultural Economics, Purdue University
rflorax@purdue.edu

Selected Paper prepared for presentation at the Agricultural & Applied

Economics Association's 2014 AAEA Annual Meeting, Minneapolis, MN, July 27-29, 2014.

The Effectiveness of International Aid: A Generalized Propensity Score Analysis

Jing Liu* Raymond Florax*

May, 2014

Abstract

Applying the generalized propensity score matching, we estimate the effect of international aid on country GDP growth based on a panel data of 73 aid recipient countries for the period 1960-2011. The matching permits us to select a set of country-year observations that to the largest extent resemble each other, thus substantially reducing the bias caused by cross-country heterogeneity. Another advantage of this analysis is that we provide the effect of aid over a continuous range of aid-receiving intensities, rather than estimating a single average effect. The does-response function appears to be a more precise description of the effectiveness of aid on development when there exists a nonlinear relationship between the two. We find a slightly concave increasing response of economic growth to international aid intensity. The optimal ratio of aid to country GDP is found to be about 6 percent, which corresponds to a 2 percent increment in country GDP growth rate.

Keywords: Propensity score matching, aid effectiveness

1 Introduction

It is intuitively expected that international aid boosts development. Yet strong and decisive empirical evidence has been slow to emerge. Recently, curiosity about the effectiveness of aid is further sparked by a series of studies that carefully documented a very normative finding - the relationship between development aid and economic growth is essentially zero (Rajan and Subramanian, 2008; Doucouliagos and Paldam, 2011). Over the past two decades, there has been an extensive and still growing effort to look for new theories, better data and advanced empirical strategies that can help understand why previous work reports mixed findings.

*Department of Agricultural Economics, Purdue University. Corresponding author: Jing Liu, 403 West State St., West Lafayette, IN 47907-2056, USA. Email: liu207@purdue.edu, Tel: (+1) 765-543-8959.

Some early reflection was ignited by the theory of conditional effectiveness emerging in the mid-1990s. It believes that aid is healthy in some cases but harmful in others, depending on certain conditions. Some commonly cited conditions include good policy, favorable institution and aid moderation. The good policy condition (Burnside and Dollar, 2000) suggests that aid works only if the recipient country pursues approperiate macro-policies (openess of trade, monetary and fiscal policies, etc.). Likewise, the institution condition argues that aid effectiveness is conditional on certain properties of a institution such as democracy, political stability and economic freedom(Collier and Dollar, 2004; Kosack, 2003). The third condition is aid itself, which suggests that aid works if given in reasonable limits, and backfires if taken in excess (Hadjimichael, 1995; Hansen and Tarp, 2000).

While some debate theories, others suspect that the difference is essentially an empirical issue. For example, Mavrotas (2005) argues that in the form of food and project aid, aid inflow tends to reduce public investment, but the opposite will be true if it comes in the form of technical assistance. Thus, solving the puzzle entails using disaggregated aid flow data. Clemens et al. (2012), however, argue that the principal reason for the failure of detecting the effectiveness of aid is that previous studies tested for aid effects within an inappropriate time horizon and looked at historical time series that were too short.

Despite the divergence, a common problem all these studies have managed to deal with is the reverse causality between international aid and economic growth. That is, economic situation can affect aid allocation, and vice versa. The most popular solution by far is the use of the Instrumental Variable (IV). But as criticized by Clemens et al. (2012), instruments that are of questionable validity and strength should be at least partially blamed for the divergent findings. Similarly, Deaton (2009) casts doubt on IV's capability of mimicking random allocation, because many of the IVs are essentially external but not plausibly exogenous.[1]

The new-generation studies, taking inspiration from the project evaluation literature, tackle this problem from different perspectives. One strategy is to conduct randomized controlled trials (RCT). The intuition behind is that the difference in the average outcome between the treated and the controls is an estimate of the average treatment effect among the treated. When the two groups are randomly selected, this effect equates to the average treatment effect for all. The good news about the RCT is that the quantitative estimation based on it does not require any modeling

[1]For example, externality indicates that population size, which is often chosen as an IV for aid, is not caused by economic growth. But exogeneity requires that population size has no effect on economic growth through the effects on aid, which makes no sense at all (Deaton, 2009). In this vein, Rubin Causal Model (RCM) (Rubin, 1978) has been used to determine causal effect by comparing potential outcomes of different levels of treatment exposure. The framework encompasses randomized experiments. However, due principally to the nature of economic data and studies, which are often not amenable to randomization, economists normally apply the framework to observational data. See Imbens and Wooldridge (2008) for a review of recent developments in the econometrics of program evaluation.

and rests on only minimal assumptions. But it is generally expensive to conduct. Moreover, the RCT is informative only about the mean of the treatment effect but not other features of the distribution (Deaton, 2009).

Doubly robust estimator is another technique to estimate treatment effect with a causal interpretation from a non-random sample (Robins and Rotnitzky, 1995). It combines propensity score model with the outcome regression model, such that any bias caused by mismodeling can be removed as long as at least one of the two models is correctly specified. This estimator does not eliminate the need for correct model specification, but provides a second chance to satisfy it. For this reason, it is described as the "best practice" by Imbens and Wooldridge (2008). However, the central difficulty in this approach is that the "either-or" assumption might not hold due to some unmeasured confounders, which is not uncommon in observational studies.

The objective of this paper, similar to that of many previous studies, is to examine the effectiveness of international aid. However, what distinguishes the present study from the others is that instead of analyzing whether or not on average receiving aid fosters development at all, we focus on to what extent a higher treatment intensity yields stronger or weaker effects than a lower treatment intensity does. Ideally, answering the question would involve an experimental approach that randomly assigns a low or high level of development aid to countries and compares their economic performance afterwards to see how effective the aid is. Such an experiment, however, would be practically difficult. As an alternative, we first balance the treatment and control groups (countries who receive more aid and those who received less aid) to eliminate bias associated with difference in the covariates. Next, the effect of aid on development is estimated using a dose-response function, which provides a predicted outcome for every level of the treatment conditional on a balanced distribution of observed covariates.

Our study contributes to the aid effectiveness literature in several ways. First of all, we advance the understanding of *causal* effect of international aid on development by using a technique that minimizes selection noise and provides estimates less likely to be biased. Second, rather than identifying the causal relationship at the mean, we aim to present the distribution of the outcome over the entire range of treatment levels.

The rest of the paper is organized as follows. Section two describes the data. Section three outlines the generalized propensity score method and the implementation procedures. Section four discusses the results from the empirical analysis.

2 Data

2.1 Outcome and treatment variables

We are interested in the effect of international aid (treatment) on economic development (outcome). Following the literature, we measure economic growth using annual growth rate of real GDP (PPP converted) per capita provided by the Penn World Table (7.1). The treatment, a continuous variable gauging the intensity of financial aid assistance, is represented by the ratio of aggregate net development assistance to GDP (both in current US dollars). The assistance includes all loans and grants undertaken by the official sector with the promotion of economic development as the main objective and where loans have a grant element of at least 25%. These records are obtained from the OECD International Development Statistics, Development Assistance Committee (DAC) aid statistics database. The final dataset consists of a panel of 73 developing countries that received bilateral and multilateral development aid from donor countries and international organizations during the period 1960-2007.

2.2 Control variables

The propensity score represents the probability that an agent takes treatment, in our case the ratio of aid to GDP, conditional on observed variables. The selection of these variables has been dominantly focused on the lag of recipient country's economic situation and policies. Related arguments are that aid is normally allocated based on needs, and that the lags are predetermined and therefore less susceptible to bias from confounding. However, Rajan and Subramanian (2008) criticize this specification on the grounds that the lags might be preset but still not exogenous, especially if the dependent variable or the right-hand-side variables are serially correlated. Alternatively, they suggest to explain aid flow based on donor-related characteristics, such that to the largest extent the possibility of explanatory variables picking up characteristics confounded with aid receiving status is reduced.

Inspired by these arguments, we model aid allocation to reflect considerations from both the donor and the recipient side. The main identification assumptions are that donor's choice of target countries is motivated by two reasons - history and influence, and that the disadvantaged countries are more likely to be targeted because the purpose of aid is to relieve. Following Rajan and Subramanian (2008), we use colonial link and language commonality to represent historical relationship between donors and recipients. To capture the effect of country influence, we use the size of the donor (measured by population) relative to the size of the recipient country, and its interaction with the colonial links. The interaction term is involved to account for the possibility that the donor's influence can be boosted by its own colonial tie with the recipient but weakened if the recipient country was ever colonized by other donors. As for the recipient-side characteristics,

we use GDP per capita and life expectancy for the year 1960 to characterize the disadvantage of recipient countries. We consider only the recipient's initial situation to minimize the potential confounding by contemporaneous measurements. The data source for country population, GDP per capita and life expectancy is the World Bank's WDI database. For colonial relationship, we use information provided by Rose (2004).

3 Methods

3.1 An outline of the Generalized Propensity Score matching

The main obstacle in identifying a causal relationship between treatment and outcome is the lack of randomness in aid allocation. In the context of international aid, factors like country characteristics and international relations can systematically determine the direction and magnitude of the aid flow. To restore the randomness, one commonly used method is the propensity score matching,[2] which compares the observable characteristics of units to screen out a control group that to the largest extent resembles the treated group. This method was first developed by Rosenbaum and Rubin (1983) in the setting of binary treatment. The simplest case was then extended into categorical multivalued treatment by Imbens (2000), and more recently, to the continuous treatment by Hirano and Imbens (2004), which is dubbed the generalized propensity score (GPS) matching.

For a sample of countries indexed by $i = 1, ..., N$, there are a set of outcomes $Y_i(t)$ (annual real GDP growth) for a given level of treatment $t \in T$ (the ratio of aid to GDP). In the binary case, the treatment would be $t \in \{0, 1\}$; while in the GPS framework the treatment is continuous, with the interval $[t_0, t_1]$, with $t_0 > 0$. The objective of the analysis goes beyond answering the question of whether or not aid boosts development on average. Rather, the interest is to show a dose-response function

$$\mu(t) = E[Y_i(t)], \tag{1}$$

which tells to what extent a higher treatment intensity yields stronger or weaker effects than a lower treatment intensity does.

3.2 Key assumption and property

Hirano and Imbens (2004) emphasize that the functionality of the GPS approach requires the weak unconfoundedness assumption and the balancing property to hold:

[2] See Heckman and Navarro-Lozano (2004) for a comparison of the propensity score matching and the other two popular correction methods - IV and control function

(i). *Weakness confoundedness*

$$Y(t) \perp T|X, \forall t \in T \tag{2}$$

Weakness unconfoundedness means that, after controlling for observable characteristics X, any remaining difference in treatment intensity T across units is independent of the potential outcomes $Y(t)$. The reason of referring to it as weak confoundedness is that it does not require joint independence of all potential outcomes, but requires only pairwise conditional independence of treatment with each of the potential outcomes. The random variable treatment T is assumed to be conditionally independent from the random variable outcome Y, measured at an arbitrarily chosen treatment level t.

(ii). *Balancing property*

Let $r(t, X)$ be the conditional density of the treatment given covariates. Thus, we have

$$r(t, X) = f_{T|X}(t|X) \tag{3}$$

The generalized propensity score is then defined as $R = r(T, X)$, for all $t \in T$. The balancing property requires that within the strata of $r(t, X)$, the probability of $T = t$ does not depend on the value of X. It basically says the control group should have some overlap with the treated group in terms of unit characteristics. If otherwise, for example after the matching, small countries always receive more aid than large countries, the aid allocation may not be a random process. Once this condition is satisfied, we are more confident that the GPS summarizes as much as possible information in the multi-dimensional vector X, such that within that particular GPS strata, aid allocation is random. Mathematically, this is expressed as

$$X \perp I\{T = t\}|r(t, X), \tag{4}$$

where $I(.)$ is the indicator function. Under this property, the randomness for the entire sample can be easily restored by controling for GPS stratas. Thus, a rigorous GPS analysis needs to show how well this condition is met and whether the application is appropriate and valid.

Hirano and Imbens (2004) proves that given (i) and (ii), one can eliminate bias associated with difference in the covariates in three steps: First, estimate the GPS - the conditional density of the treatment given the covariates; Second, estimate the conditional expectation of outcome $E(Y|T = t, R = r)$ as a function of the treatment and the GPS; Third, estimate the dose-response function, $\mu(t) = E[\beta(t, r(t, X))], t\forall T$, by averaging the estimated conditional expectation, $\hat{\beta}\{t, r(t, X)\}$ over the GPS at every level of the treatment we are interested in.

3.3 Implementation

Closely following Hirano and Imbens (2004), we assume a normal distribution for the treatment intensity given the covariates:

$$T_i|X_i \sim N(\beta_0 + X_i\beta_1, \sigma^2) \tag{5}$$

After obtaining the predicted level of treatment, we then estimate the GPS as follows.

$$\hat{R}_i = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}\left(T_i - \hat{T}_i\right)^2 \right) \tag{6}$$

Next, the conditional expectation of outcomes given the treatment and the GPS is specified as

$$E[Y_i|T_i, R_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 T_i^3 + \alpha_4 \hat{R}_i + \alpha_5 \hat{R}_i^2 + \alpha_6 \hat{R}_i^3 + \alpha_7 \hat{R}_i T_i \tag{7}$$

Finally, the dose-response function is estimated over the entire range of treatment

$$E[\hat{Y}_i] = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \hat{r}(t, X_i) + \hat{\alpha}_4 \hat{r}(t, X_i)^2 + \hat{\alpha}_5 t\hat{r}(t, X_i)\right) \tag{8}$$

The dose-response function describes the average impact of each level of the treatment $t \in T$ for every observed level of the GPS given the covariates.

3.4 Assessing the satisfaction of weak unconfoundedness assumption and balancing property

The weak unconfoundedness assumption, also known as the assumption of selection on observables, is an identifying assumption that is not statistically testable. As explained in the data section, we carefully choose these observables to minimize the possibility of reverse causality - one of the most commonly blamed case of violating the weak unconfoundedness assumption.

To check whether the GPS adjustment establishes a sufficient balancing of covariates to eliminate the selection bias, we adopt the "blocking the score" approach proposed by Hirano and Imbens (2004). The whole sample is first divided into four groups, according to treatment quartiles. For each group, we find the median treatment value T_M^j, $j \in \{1, 2, 3, 4\}$, and evaluate $\hat{R}_i(T_M^j, X_j)$ for each unit at its own observables and T_M^j, $j \in \{1, 2, 3, 4\}$. Group j will serve as the treated group if the GPS is evaluated at T_M^j.

The next task is to use $\hat{R}_i(T_M^j, X_j)$ to screen out a control group from the pool of units that do not belong to group j. As a start of the block matching, we break group j into five blocks, according to the GPS quintiles $\hat{R}_i(T_M^j, X_j)$. It is important to note that only scores in the treated group

are involved in determining the cutting points of GPS blocks. Then for each unit $\notin j^{th}$ group, we assign it into the corresponding block depending on which interval the unit's GPS falls into. Since propensity score reflects the probability of an agent being selected for treatment, units assigned to the corresponding control group in theory have a similar chance of being treated, although they are actually not treated. Hence, they represent a counterfactual comparable counterpart of the treated group. By definition, the sample size is roughly the same across treated GPS blocks, but not necessarily so (perhaps most often not) across control GPS blocks. Table 2 summarizes the number of observations in each treated and control group.

The matching procedure prepares us to test whether within each GPS interval the means of the covariates in the treatment group are significantly different from those in the control group. For instance, when group 1 is used as a treated group, we compare the units in {group 1; block 1} with the units in {group $2, 3, 4$; block 1} and obtain the difference in two means. Altogether we calculate five sets of difference in means (one for each GPS interval). The final sample mean for the one-sample two-tail student t-test is constructed as the weighted average of the five, where the weight is the sample size of each treated and control pair combined. The standard deviation is obtained by bootstrapping. The same procedure is then repeated for another three times, alternately using group $2, 3, 4$ as treated group. If the test fails to reject the hypothesis that the two sets of units are statistically indifferent, we have more confidence to believe the treated and control groups matched by the GPS are balanced. By implication, the treatment effect obtained by comparing such a balanced pair is less likely to be biased.

4 Empirical results

4.1 Estimate the GPS

We estimate Equation (4) via the least square estimator with year and continent indicators and report the results in Table 1. To allow for a non-linear relationship between the intensity of treatment and initial GDP per capita, we include a quadratic and a cubic of log GDP per capita along with the linear effect. Almost all the covariates exert a significant impact on treatment intensity with the expected signs.[May need a little more interpretation of the results here.]

The negative coefficients for the interaction of donor's influence with donor-specific dummies suggest that the recipient's tie with the most powerful colonizer weakens the aggregate influence from all the other donors, or a "scare-away" effect that prevents others from seeking influence. Based on these estimates, the GPS for each unit is estimated using the formula given by Equation (5). The estimated GPS along with the actual treatment intensity received by each unit will then be used in the second step to estimate the conditional expectation of the outcome Y_i given T_i and $\hat{R}_i$.

4.2 Check balancing property

As a benchmark for comparison with the balancing after propensity score matching, we test in the pre-matching sample whether the mean in one treatment group is different from the mean in the other three treatment groups combined. In the left panel of Figure 1, we plot the t-values derived from the Student-t test. The result shows little sign of balance, with only 9 of the 56 t-values have absolute values of less than 1.96, indicating the "treated" group and the "controlled" group differ starkly in the observables. Next, we employ the matching algorithm described in section 3.4 and re-plot the t-values, as shown in the right panel of Figure 1. The result shows a significant improvement in balancing. 42 out of 56 t-values have absolute values of less than 1.96, suggesting the "treated" and the "controlled" group are much more silimar to each other.

4.3 Estimating the dose-response function

We estimate equation (6) via least square estimator and predict the expected country GDP growth rate given propensity score and aid intensity. Results are reported in Table 3. It's important to note that $\alpha(t, r)$ parameters do not have a causal interpretation. They are estimated however for two reasons - to generate the dose-response function and to check whether the non-randomness matters. If it does matter, we expect the GPS-related parameters to be statistically significant either individually or jointly. When that happens, propensity score matching is relevant and preferable in the sense that it helps tease out the causal relations between aid and economic growth.

Finally, the average impact of aid, as formulated in Equation (7), is plotted against the entire range of aid distribution in Figure 2. The figure reveals a nonlinear relationship between international aid and economic growth. As one country receives more aid relative to its GDP, the economy keeps growing until the treatment reaches a critical level, beyond which more intensive assistance would be harmful. In our sample, this tipping point takes place at 6 percent within the interval of 2.36×10^{-7} to 0.14. The GDP growth rate corresponds to this aid intensity is 2.15 percent. Furthermore, the magnitude of the marginal treatment effect varies, depending on aid intensity.

5 Conclusion

Applying the generalized propensity score matching, we estimate the effect of international aid on country GDP growth based on a panel data of 73 aid recipient countries for the period 1960-2011. The matching permits us to select a set of country-year observations that to the largest extent resemble each other, thus substantially reducing the bias caused by cross-country heterogeneity. Another advantage of this analysis is that we provide the effect of aid over a continuous range of aid-receiving intensities, rather than estimating a single average effect. The does-response function

appears to be a more precise description of the effectiveness of aid on development when there exists a nonlinear relationship between the two. We find a slightly concave increasing response of economic growth to international aid intensity. The optimal ratio of aid to country GDP is found to be about 6 percent, which corresponds to a 2 percent increment in country GDP growth rate.

Table 1: Estimation of the generalized propensity score (GPS).

Variable	Coef.	Std. Err.
Intercept	-15.69	3.20
$\log(GDPPC)$	9.83	1.57
$(\log(GDPPC))^2$	-1.68	0.28
$(\log(GDPPC))^3$	0.08	0.02
Initial life expectancy	-1.43	0.22
Ever had a colonial relationship, if = 1	-4.01	0.49
Proportion of language-sharing donors	0.06	0.24
UK colony, if =1	3.65	0.37
France colony, if =1	2.74	0.51
Spain colony, if =1	4.81	0.54
Ratio of log of pop. of donor/recipient	0.41	0.33
Ratio of log of pop. of donor/recipient × Ever coloized dummy	4.26	0.45
Ratio of log of pop. of donor/recipient × UK colony dummy	-3.61	0.33
Ratio of log of pop. of donor/recipient × France colony dummy	-2.94	0.45
Ratio of log of pop. of donor/recipient × Spain colony dummy	-4.51	0.48
Observations	3462	
R^2	0.584	
Year dummy	Yes	
Continent dummy	Yes	

Table 2: Number of observations in each treated and control group

Block	Treated 1	Control 1	Treated 2	Control 2	Treated 3	Control 3	Treated 4	Control 4
1	173	1731	173	1405	173	878	173	1238
2	173	392	173	402	173	455	173	244
3	173	106	174	327	174	308	173	111
4	173	77	173	227	173	233	173	63
5	173	48	173	180	173	219	173	50

Table 3: Estimation of the dose-response function.

Variable	Coef.	Std. Err.
$Intercept$	2.23	0.83
Aid/GDP	14.01	27.26
$(Aid/GDP)^2$	-143.22	176.20
$(Aid/GDP)^3$	258.68	271.15
$Score$	3.38	20.36
$Score^2$	1.72	137.62
$Score^3$	-36.55	264.85
$Score \times (Aid/GDP)$	-189.86	78.99
Observations	3166	
R^2	0.010	
Year dummy	No	
Continent dummy	No	

Figure 1: Balance of covariate accounting for the GPS.

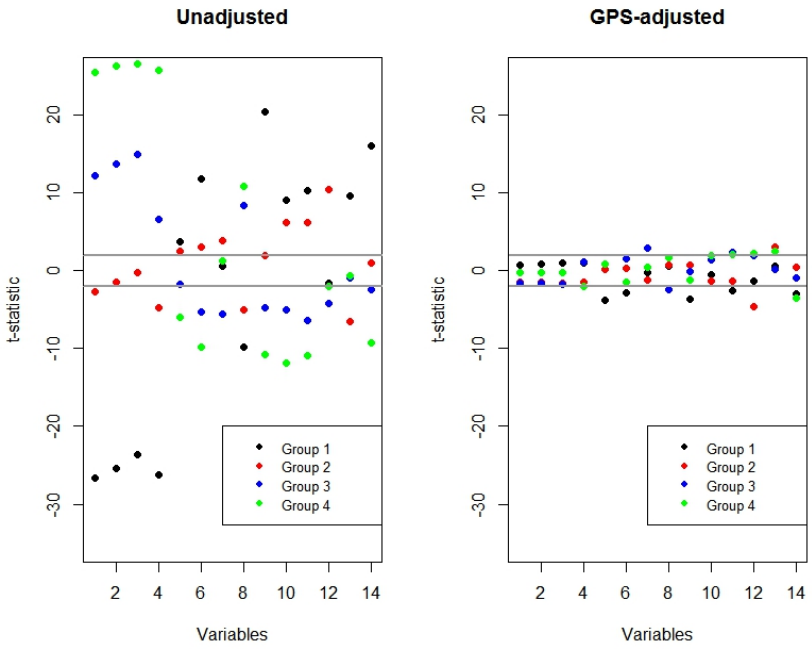
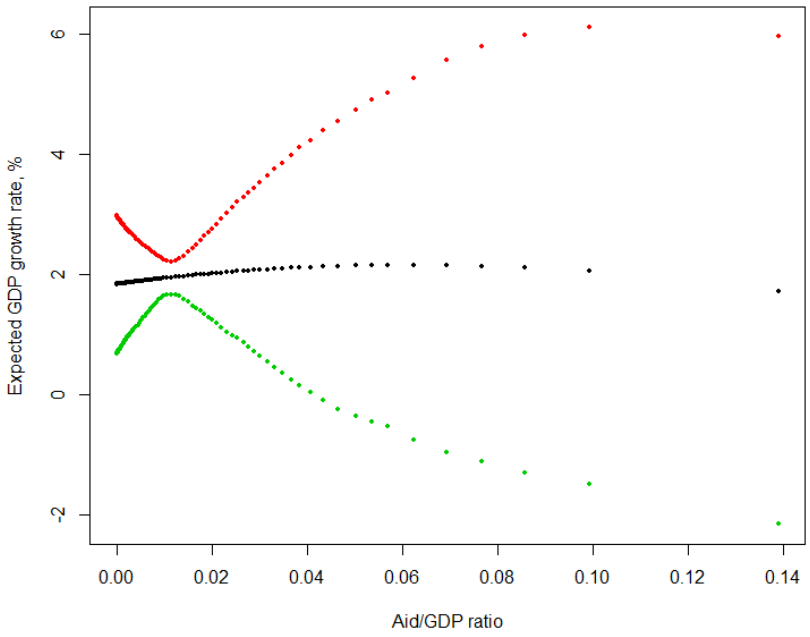


Figure 2: Dose-Response Function



References

Burnside, C. and D. Dollar (2000). Aid, policies, and growth. *American economic review*, 847–868.

Clemens, M. A., S. Radelet, R. R. Bhavnani, and S. Bazzi (2012). Counting chickens when they hatch: Timing and the effects of aid on growth*. *The Economic Journal 122*(561), 590–617.

Collier, P. and D. Dollar (2004). Development effectiveness: what have we learnt?*. *The economic journal 114*(496), F244–F271.

Deaton, A. S. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Technical report, National Bureau of Economic Research.

Doucouliagos, H. and M. Paldam (2011). The ineffectiveness of development aid on growth: an update. *European journal of political economy 27*(2), 399–404.

Hadjimichael, M. T. (1995). *Sub-Saharan Africa: growth, savings, and investment, 1986-93*, Volume 118. International Monetary Fund.

Hansen, H. and F. Tarp (2000). Policy arena aid effectiveness disputed. *Journal of International development 12*(3), 375–398.

Heckman, J. and S. Navarro-Lozano (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and statistics 86*(1), 30–57.

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives 226164*, 73–84.

Imbens, G. M. and J. M. Wooldridge (2008). Recent developments in the econometrics of program evaluation. Technical report, National Bureau of Economic Research.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Kosack, S. (2003). Effective aid: How democracy allows development aid to improve the quality of life. *World Development 31*(1), 1–22.

Mavrotas, G. (2005). Aid heterogeneity: looking at aid effectiveness from a different angle. *Journal of International Development 17*(8), 1019–1036.

Rajan, R. G. and A. Subramanian (2008). Aid and growth: What does the cross-country evidence really show? *The Review of economics and Statistics 90*(4), 643–665.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–129.

Rose, A. K. (2004). Do we really know that the wto increases trade? *The American Economic Review 94*(1), 98–114.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34–58.