HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

THE UNIVERSITY of York

HEDG Working Paper 11/02

# Using a Semiparametric Estimator to Forecast Education Outcomes in Nicaragua's Red de Proteccion Social

**Ranjeeta Thomas**

**February 2011**

york.ac.uk/res/herc/hedgwp

# Using a Semiparametric Estimator to Forecast Education Outcomes in Nicaragua's Red de Protección Social

Ranjeeta Thomas *

February 28, 2011

**Abstract**

This paper uses baseline data from the randomized experiment of the conditional cash transfer program - Red de Protección Social, Nicaragua to conduct an *ex ante* evaluation and compares results to those of the experimental evaluation. Reduced form estimation of a behavioural model using a health production framework forms the basis of the evaluation. A Klein and Spady semi-parametric single index model is used to predict unobserved outcomes under the treatment. The sample consists of children aged 7-13 who have not completed grade 4 of primary school. The evaluation shows that the *ex ante* approach closely matches the experimental outcomes in most cases.
*ex ante* evaluation conditional cash transfers single index models education

**Keywords:***ex ante* evaluation, conditional cash transfers, single index models, education, health production function

**JEL Classification** I38 · I20 · I12 · O12

## 1 Introduction

The last few decades has seen billions of dollars channelled to developing countries as international aid. Despite this impetus these regions continue to remain amongst the poorest in the world with some of the worst indicators of poverty and health. Improving the critical link between aid and outcomes requires ensuring resources are channelled to where they are likely to have the greatest impact (White, 2006). Research on this link involves the evaluation of development programs to measure their impact. Most of this research has focussed on ex post evaluations of programs by either randomized-experimental allocation of the

*I would like to thank Prof. Andrew M. Jones, Prof. Nigel Rice and the members of the Health Econometrics and Data Group (HEDG) at the University of York, UK for their comments. This paper uses data made available by the International Food Policy Research Institute (IFPRI). All errors remain my responsibility. Thomas: University of York, Department of Economics and Related Studies, Heslington, York YO105DD, rat503@york.ac.uk

intervention or using observational approaches such as difference-in-differences, matching methods or regression discontinuity. On the other hand, applications of forecasting in economics have been widely applied in estimating demand and predicting impacts of macro-economic policies but there are comparatively few applications in evaluating social programs. In development settings with constraints on resources, ex ante evaluations are particularly useful in making informed decisions for extending the target population of an existing program. They also facilitate optimal usage of limited resources by ensuring governments make financial investments in programs that are likely to have a positive impact. These evaluations are useful in considering implementation of new programs and serve as complements to future ex post evaluations.

Leading examples of *ex ante* evaluations of a social program are Todd and Wolpin (2006) and Attanasio et al (2005), who evaluate the impact of Mexico's Progresa Conditional Cash Transfer program by structurally estimating the parameters of a behavioural model that specifies the interactions of the program. In contrast to the structural estimation approach in a recent simpler reformulation, Todd and Wolpin (2010) build on the work by Ichimura and Taber (2000) and illustrate the use of reduced form estimation of behavioural models in evaluating social programs without specification of functional forms. The authors illustrate situations in which a non-parametric estimation strategy based on a behavioural model can be used to estimate ex ante impacts. This reduced form *ex ante* approach differs from *ex post* evaluations in the way it uses the traditional potential outcomes framework in that the data are observed for only the untreated population. In this case the counterfactual to be estimated are the outcomes for the population when treated rather than for the controls. Program impacts using the behavioural model reduced form (BMRF) approach are estimated from an underlying economic model and use variation in the policy variable for model identification.

The objective of this paper is to apply the reduced form estimation approach to a program that focuses on education and health and to compare the predicted outcomes with results from a randomized experiment. It is based on an economic model of household consumption and uses data from the experimental evaluation of Nicaragua's Red de Protección Social (RPS), a conditional cash transfer (CCT) for rural households in Nicaragua. The program aims at improving school enrolment and attendance of children age 7-13 who have not completed grade 4 and health and nutritional status of children below 5 years by supplementing household income through the cash transfers. The cash transfers are conditional on a certain minimum school attendance by children of recipient households and attendance at health workshops by mothers.

This paper extends the approach applied by Todd and Wolpin (2010) by applying it to the RPS program involving a model that can jointly determine school and health outcomes. The estimation strategy uses variation in the costs of schooling, full wealth of households along with several household characteristics

to determine the impact of the program. The large number of covariates determining outcomes does not permit fully non-parametric estimation. To overcome this, a semi-parametric single index model for binary outcomes is used to predict impact. This paper presents the economic model estimates the impact on school enrollment of the program estimated using data from the randomized experiment.

Section 2 of the paper describes the program in detail listing out the goals, conditionalities and eligibility criteria, Section 3 provides the underlying model, Section 4 provides the empirical strategy, Section 5 presents the data and Section 6 discusses results.

The key identification condition in this approach is that the program has an impact only through the budget constraint of the behavioural model (**?**), ensuring that the reduced form before the program is also the same after, except for a change in the magnitude of the variables resulting from the program. The approach also relies strongly on selection on observables to capture heterogeneity. As specified by **?** extending the approach to allow impact of the program to affect preferences would require specification of some functional form. In such cases stronger assumptions are required and the similarity of the reduced form before and after the program will depend on the nature of the functional form assumed.

# 2 Red de Protección Social

*Red de Protección Social* (RPS) is based on the design of Mexico's PROGRESA program and is the first CCT to be implemented in a low-income country[1]. The program was introduced in 2000 and targets reducing financial barriers to accessing education and health care in rural Nicaragua. In 1998 data from the Living Standards Measurement Survey (LSMS) indicated that 48% of the population in Nicaragua was poor and 75% of this population lived in rural areas (World Bank 1998). The program was implemented in two phases. The first phase was designed as a pilot randomized experimental evaluation in two districts, Madriz and Matagalpa, based on the level of poverty and capacity of these districts to implement the program. In both these regions 80% of the rural population was poor and of this population 50% were exteremely poor (Maluccio and Flores, 2005; IFPRI, 2001a). Phase Two of the project extended the program for a futher 3 years. This paper uses data from the pilot phase of the program generated by the randomized experiment. From the two regions selected 42 *comarcas* or administrative units were selected based on a marginality index to participate in the pilot.

[1]This section is based on the description of RPS provided by Maluccio and Flores (2005) in the impact evaluation report of the randomized experiment

The program consisted of two demand side components - the first focusing on food security, health and nutrition and the second on education. Each eligible household received a "food security transfer" every alternate month based on two conditions, attendance at educational workshops held every other month and children under age 5 being brought for scheduled preventive health check-ups. The demand side health initiatives were complemented by supply-side enhancements including training and payment to private health care providers to ensure the increased demand from the program was met. The food transfer was a fixed amount and did not depend on the size of the family. The education component of the program consisted of two cash transfer components to families with children aged 7-13 who had not completed grade 4 of primary school, conditional on enrolment and regular attendance by the children. The first was a lump-sum transfer provided as a fixed amount per family regardless of the number of eligible children, conditional on all eligible children enrolling in school. In addition a cash transfer for school supplies was provided for each eligible child, also conditional on enrolment. On the supply-side, incentives were provided to teachers to compensate for the additional monitoring and reporting required to ensure compliance with the program and the increase in class size from the enrolments. Figure 1 presents a summary of the eligibility criteria and requirements for RPS.

The transfers target a reduction in the net price of schooling and food consumption to reduce short-term poverty while encouraging investments in human capital to eliminate long-term poverty. The amounts of the transfer include the Córdoba 2000 equivalent of US$224 for food security and US$112 for the educational component. The school supplies component for each eligible family was US$21. Figure 2 presents a summary of the transfers. According to the *ex post* evaluation of RPS by Maluccio and Flores (2005) the food transfer was equivalent to 13% of annual household expenditures and families with one eligible child for the schooling components would receive an additional 8% of annual household expenditures. Beneficiaries that did not comply with the specific requirements associated with each component failed to receive the transfer for the particular component.

The randomized evaluation provide census data (for all eligible households and individuals) in the 42 selected *comarcas*, baseline data for the final selection of households based on the marginality index after assignment into treatment and control groups and follow-up data for the next two years. Since the objective of this paper is an *ex ante* evaluation, the focus is on data generated prior to the introduction of the intervention.

# 3 Economic Framework

The model uses the household production framework of Becker (1965). A household with multiple eligible children $i = 1...n$, has utility $U$ a function of $C$ representing non-medical consumption, health status of each child $H$, a binary indicator of school enrolment $S$, with $S = 1$ indicating school enrolment, and an indicator of gender $g$. The household maximisation problem is then:

$$\max_{C,S} U(C, S_i; g) \tag{1}$$

The time constraint for an eligible child can be written as:

$$T_i = T_{si}.S_i + T_{wi}(1 - S_i) \tag{2}$$

where $T_{si}$ is time spent in school and is assumed to be a fixed amount for all enrolled children, $T_{wi}$ is time spent at work.

The money budget constraint can be written as:

$$C + \sum_{i=1}^{n} \delta_i.S_i + p_m.M - \mu.E^m = Y + \sum_{i=1}^{n} w.T_{wi}(1 - S_i) \tag{3}$$

Where $\mu.E^m = 0$ in the pre-program scenario, $\delta_i$ is the direct cost of schooling for child $i$. Primary schooling in free in Nicaragua and most children face no tuition fees, hence $\delta_i$ includes all other school related costs faced by families such as transport, uniforms, books and school meals. $p_m$ is the cost per unit of medical care consumed and $Y$ is household income net of the earnings of the program eligible children. With $\mu.E^m = 0$ in the pre-program scenario the money budget constraint is:

$$C + \sum_{i=1}^{n} \delta_i.S_i + p_m.M = Y + \sum_{i=1}^{n} w.T_{wi}(1 - S_i) \tag{4}$$

The constrained household maximisation problem is:

$$\max_{C,S,T_w,T_s} U(C, S_i; g) \tag{5}$$

The full income constraint combining both the time and money constraint is:

$$C + \sum_{i=1}^{n} [\delta_i + w.T_{si}]S_i + p_m.M = Y + w.\sum_{i=1}^{n} T_i = F \tag{6}$$

where $F$ is full income of the household and the total price of schooling for all eligible children in the family ($\theta = \sum_{i=1}^{n} [\delta_i + w.T_{si}]$) is the cost of schooling plus the shadow wage for the eligible children.
The optimal choice of schooling is $S^* = \Phi(F, \theta, p_m, n; g, X_h)$

The RPS program has two cash transfers - the first focuses on changing the price of schooling for eligible children conditional on enrolment and the second is a food transfer meant to boost consumption, nutrition and access to preventive health care conditional on mothers' attending the health workshops. The initial objective of RPS was to condition the food transfer on a series of other requirements including taking children under 5 years for health checks and maintaining up-to-date immunization. But as explained in the program description this conditionality was not enforced till almost the second year of the program and hence does not affect the analysis in this paper. The household food transfer ($\mu$) conditional on $E_m$ is modelled as a direct income effect, raising the income level of the household and does not stipulate specific expenditure categories. At subsistence consumption levels, an increase in income through a transfer is assumed to impact food consumption changing consumption patterns to more nutritious components in the food basket and reducing financial barriers to utilizing preventive care.

The school transfer is implemented as two components $(\tau, \rho)$ to reduce the net price of schooling and substitute for any wages earned by children not enrolled in school due to employment. Schooling and labour market participation are assumed in the model to be substitutes. A decrease in the price of schooling is likely to encourage children to substitute away from labour market participation and increase school enrolment. The first component $\tau$ is provided for each eligible child in the family while $\rho$ is a lump sum transfer irrespective of the number of eligible children. Both transfers are conditional on all eligible children enrolling in school.

With the introduction of the subsidies $\mu.E^m$, $\tau.\sum_{i=1}^{n} S_i.S_p$ and $\rho.S_p$, where $S_p = 1$ if $\sum_{i=1}^{n} S_i = n$ ie. all eligible children enrol in school and $S_p = 0$ otherwise. The money budget constraint can be written as:

$$C + \sum_{i=1}^{n} \delta_i.S_i + p_m.M = Y + \sum_{i=1}^{n} w.T_{wi}(1 - S_i) + \mu.E^m + \sum_{i}^{n} \tau.S_i.S_p + \rho.S_p \tag{7}$$

The full income constraint is then:

$$C - \rho.S_p - \mu.E^m + \sum_{i=1}^{n} (\delta_i + w.T_{si} - \tau.S_p) S_i + p_m.M = Y + \sum_{i=1}^{n} w.T_i = \tilde{F} \tag{8}$$

The new price of schooling under the subsidy program is $\tilde{\theta} = (\sum_{i=1}^{n}[\delta_i + w.T_s - \tau.S_p])$ and the cost of consumption is $C - \rho.S_p - \mu.E^m$. The optimal choice under the subsidies is $S^{**} = \Phi(\tilde{F}, \tilde{\theta}, p_m, n; g, X_h)$ and health is $H^{**} = \Omega(\tilde{F}, \tilde{\theta}, p_m, n; g, X_h)$

Empirically this allows exploitation of two sources of variation in the data to compare untreated individuals with outcomes $S^*$ with other untreated individuals with outcomes $S^{**}$ - the first is school costs and the second is full income of the households at the baseline. As described earlier, primary education is free in Nicaragua and most families face no fees, the cost here includes other expenditure related to schooling which is exogenous in the sense that it is faced by all families when enrolling children irrespective of whether the tuition is free or not. Figure 3(a) shows a histogram of full income of a families, with values ranging from c1,590 to c77,905. The second graph figure 3(b) shows the school costs used in the estimation range from c12 to c1438. In addition to variation in school costs and full income, the level of the school grant also varies depending on the number of children in the household. The treatment effect is estimated by matching the treated and untreated groups on functions of observable characteristics. Identifying the *ex ante* treatment effect also requires that any unobserved heterogeneity ($\nu$) remains the same before and after treatment ie. ($\nu$) is independent of full income and school costs. To make this assumption plausible, empirically the matching functions include a set of family characteristics.

$$f(\nu|F, \theta) = f(\nu|\tilde{F}, \tilde{\theta}, X_h)$$

# 4 Empirical Specification

The above approach generates a set of variables that naturally extend to an empirical application of the model. This relies on direct variations in the policy variables. In this case variation in the costs of schooling and health care can be exploited and a matching estimator applied to identify predicted program impacts.

Typical evaluation exercises using information on treated outcomes ($S_1$) estimate the counterfactual of untreated outcomes ($S_0$). In contrast, in the *ex ante* approach treated outcomes are unobserved and are the counterfactual that needs to be estimated. From the model, as indicated by Todd and Wolpin (2010), the unobserved $S_1$ can be represented in terms of the observed untreated outcomes conditional on an equivalent set of exogenous variables. This idea can be represented as:

$$S_{1i} = E[S_{0j}|F_i = \widetilde{F_j}, \theta_i = \widetilde{\theta_j}, n_i = n_j, g_i = g_j, X_{hi} = X_{hj}] + \epsilon \tag{9}$$

Todd and Wolpin propose a matching estimator of the average treatment effect for those eligible for the program (intent-to-treat (ITT)) as:

$$\alpha = \frac{1}{k}\sum_{j=1}^{k}{}_{j,i\in S_p}E(S_i|F_i=\widetilde{F_j},\theta_i=\widetilde{\theta}_j,n_i=n_j,g_i=g_j,X_{hi}=X_{hj}) - S_j(F_j,\theta_j,n_j,g_j,X_{hj}) \qquad (10)$$

## 4.1 Estimating School Costs

Implementing the above matching estimator requires estimation of the unobserved treated outcomes as a function of household expenditure, school costs, medical care expenditure and a set of household characteristics. School costs ($\delta_i$) are determined by the enrolment status of the child and hence are observed in the data for only those children who are currently enrolled in school and zero costs observed for those not enrolled. The problem of predicting school costs for the entire sample of children requires using a two-step process decomposing the participation decision and the determinants of the cost of schooling. A *two-part model (2PM)* is applied where in the first part, the enrolment decision, is modelled using a probit and the second part predicts the cost of schooling as a linear function of the determinants of school costs (Mullahy, 1998). The most common specification of the second part is a log transformation of the outcome variable. A problem with using a retransformed OLS in this case is that zero school costs are also observed in the sample of those children currently attending school. A log transformation would drop these observations from the estimation sample. A further problem arises with retransformation of the outcome variable to the original scale in the presence of heteroskedasticity. Manning (1998) shows that heteroskedasticity leads to biased estimates of the outcome variable and correction requires determining whether the heteroskedasticity is across different groups or caused by a particular subset of the covariates. To overcome these issues the second part of the *2PM* is estimated using the *extended estimating equations model (EEE)* proposed by Basu and Rathouz (2005). The EEE approach is an extension of a standard *generalized linear model (GLM)* incorporating flexible link and variance functions. Specifically, the EEE combines a Box-Cox transformation for the link function and includes a class of link functions represented by an estimated parameter $\lambda$:

$$\frac{\mu^{\lambda}-1}{\lambda}$$

It also allows for heteroskedasticity and uses a general power function for the variance defined by two-parameters $\theta_1$ and $\theta_2$:

$$\theta_1 \mu^{\theta_2}$$

The model is estimated separately for boys and girls.

## 4.2 Estimating Counterfactual Outcomes

The unobserved outcomes $E(S_i|F_i = \widetilde{F_j}, \theta_i = \widetilde{\theta_j}, N_i = N_j, g_i = g_j, X_{hi} = X_{hj})$ can be estimated using a binary response model to estimate the conditional probability $P(S = 1|X = x) = G(x\beta)$. If the distribution function $G$ is known a priori then a parametric specification such as a logit or probit can be used. Misspecification of $G$ would however result in inconsistent estimates of $\beta$ and inaccurate predictions of the unobserved outcomes. To increase the flexibility and avoid misspecification problems the unobserved outcomes are estimated by regressing current enrolment status on income, estimated school costs, medical care expenditure and a set of family and child characteristics to capture unobserved heterogeneity using a semiparametric *single-index model*. The single-index model defines the conditional mean function as:

$$E(Y|x) = G(x\beta) \tag{11}$$

where $\beta$ is an unknown vector and $G$ is an unknown function and $x\beta$ represents an index. The above index specification could be made entirely flexible using a fully nonparametric approach to model outcomes eliminating the risk of any misspecification. Such an approach is however constrained in this case by the dimensionality of the covariate vector $(x)$. Nonparametric approaches suffer from the *curse of dimensionality* where convergence rates are inversely related to the number of continuous covariates and tend to be less precise as the dimension increases. The single-index $x\beta$ reduces the dimensionality problem by aggregating across $x$ and has the same convergence rate as a single dimensional quantity represented by $x\beta$. The single-index model also has advantages for predictions as the region of support extends beyond the observed $x$ to points not in the support of $x$ but in the support of $x\beta$ (Horowitz, 1998). However, unlike the nonparametric approach it builds in a parametric assumption of the linearity of the index.

The single-index model involves the joint estimation of the two unknown elements $\beta$ and $G$. Estimation of both elements require several identification restrictions. Similar to all linear models, identification of $\beta$ requires $G$ to be a non-constant function along with the absence of multicollinearity amongst the covariates. In addition, to uniquely identify the function $G(x\beta)$ single-index models involve *location normalization* and *scale normalization* restrictions. Location normalization is achieved by requiring the covariate vector to

include no intercept term while scale normalization involves restricting the $\beta$ coefficient of one continuous variable to equal one. Identification in single-index models is achieved because the conditional mean function can remain constant with changes in $x$ as long as the index $x\beta$ remains constant. However, with continuous covariates a constant index (ie. $x\beta = k$) for a given set of covariates has probability zero. To overcome this a further identification restriction is required where $G$ is a differentiable function so that $G(x\beta)$ is close to *G(k)* when $x\beta$ is close to $k$ (Horowitz, 1998). A final set of restrictions are required when $X$ contains both discrete and continuous variables. The first of these requires that the discrete elements of the covariate vector do not divide the support of $x\beta$ into disjoint subsets. The final restriction is referred to as the 'non-periodicity condition' for the function $G$ requiring it to be strictly increasing.

The single-index model defined in (13) was adapted to binary outcomes by Klein and Spady (1993). In the case of binary outcomes such as enrolment (where $S = 0, 1$) the index function is defined as:

$$E(S|x) = P(S = 1|x) = G(x\beta)$$

In a parametric setting with known $G$, $\beta$ could be estimated efficiently using a maximum likelihood estimator(MLE) where the log-likelihood is:

$$\ln L(\beta, G) = n^{-1} \sum_{i=1}^{n} [S_i \ln G(x_i\beta) + (1 - S_i) \ln(1 - G(x_i\beta))] \tag{12}$$

In the semiparametric case following Ichimura (1993), Klein and Spady propose to estimate $\beta$ by maximising the (quasi) log-likelihood function(14) replacing the unknown function $G$ with a semiparametric likelihood estimate $G_n(x_i\beta)$. 'The index restriction permits multiplicative heteroskedasticity of a general but known form and heteroskedasticity of an unknown form if it depends only on the index' Klein and Spady (1993). $G_n$ is estimated using a leave-one-out nonparametric estimator of the density of $x\widehat{\beta}$ conditional on $S$, where for any $z$

$$G_n(x_i\beta) = \frac{P_n g_n(z|S = 1)}{P_n g_n(z|S = 1) + (1 - P_n) g_n(z|S = 0)} \tag{13}$$

where $g_n$ is the kernel estimate of the conditional density of $x\beta$ ($g(.|S)$) and $g_n$ is defined as:

$$g_n(z|S = 1) = \frac{\sum_{i=1}^{n} S_i K(z - x_i\widehat{\beta})/h_n}{nP_n h_n} \tag{14}$$

$$g_n(z|S=0) = \frac{\sum_{i=1}^{n}(1-S_i)K(z-x_i\widehat{\beta})/h_n}{n(1-P_n)h_n} \tag{15}$$

where $P_n$ is the empirical probability $P_n = \sum_{i=1}^{n} S_i$, the proportion of children currently enrolled in school, $K$ is a kernel function and $h_n$ is the bandwidth.

Klein and Spady show that the estimator is asymptotically efficient and achieves the semiparametric efficiency bounds of Chamberlain (1986) and Cosslett (1987). The resulting vector of parameter estimates $(\hat{\beta})$ is shown to have the following properties:

$$n^{1/2}(\widehat{\beta} - \beta) \longrightarrow_d N(0, \Omega)$$

$$\Omega = E\left\{\left[\frac{\partial G(X_i\beta)}{\partial \beta}\right]\left[\frac{\partial G(X_i\beta)}{\partial \beta}\right]^T\left[\frac{1}{G(X_i\beta)(1-G(X_i\beta))}\right]\right\}^{-1}$$

To estimate the treatment effect (intent-to-treat) for the conditional cash transfer program on enrolment the two groups must be matched on a set of observable covariates. Identification using the estimator in 10 assumes that selection into treatment and control is solely on the basis of observable characteristics. This assumes that the distribution of any unobserved heterogeneity$(\nu)$ remains the same before and after treatment ie. $(\nu)$ is independent of income, school costs and health care costs. To make this assumption plausible the matching conditions on a set of family characteristics:

$$f(\nu|F, \theta, X_h) = f(\nu|\tilde{F}, \tilde{\theta}, X_h)$$

The unobserved treated outcomes $E(S_i|F_i = \widetilde{F_j}, \theta_i = \widetilde{\theta_j}, N_i = N_j, g_i = g_j, X_{hi} = X_{hj})$ are estimated using the Klein and Spady estimator by regressing school enrolment status of the observed control group children on observed income, school costs, age, number of children eligible for the program, number of children of under 5 years, education of the household head, health expenditure. Scale normalization is achieved by setting the coefficient for number of children under 5 equal to 1. The estimated model is then used to first predict enrolment outcomes for the observed control group observations and then extrapolate the predictions under treatment by evaluating the function at $(\tilde{F}, \tilde{\theta}, \tilde{\lambda})$.[2]

Both within sample predictions and extrapolation can only be carried out in regions of common support.

[2]The statistical package np (Hayfield and Racine, 2008) available for the software R was used. The model was run separately for boys and girls. The scalar bandwidth for the index $x\beta$ for boys is 0.083 and for girls is 0.065.

In the original formulation of the model, as required by the QMLE asymptotic theory, Klein and Spady introduce trimming procedures on the likelihood function (14) to ensure that $G$ is bounded away from 0 and 1. But their simulations show that trimming has little impact in empirical applications. Following their findings and other applications of this model (Horowitz, 1993; Gerfin, 1996; Fernández and Rodríguez-Poo, 1997) the likelihood function is not trimmed before predicting outcomes for the observed data. Extrapolation in nonparametric models is only valid at points with positive data density. The region of support $S_p$ is defined as $S_p = (x\beta) \in R^2$ such that $f(x\beta) \geq 0$ where $f(x\beta)$ is the nonparmetric density of the linear index[3]. Heckman et al (1997) propose that the density should be strictly positive as defined by $S_p$ and should exceed a minimum cut-off to avoid points with very low density. Thus the extrapolation is valid for only those points of evaluation where

$$f(x\hat{\beta}) > c \tag{16}$$

Heckman et al (1997) recommend setting the cut-off at a percent quantile of the estimated densities. Here $c$ is set at the 2% quantile. Only those observations that meet the above criterion are kept in the extrapolation sample.

# 5 Data and Variables

This paper uses data collected for the *ex post* randomized evaluation of RPS. Two datasets (IFPRI, 2005) from the *ex post* evaluation are applicable, the first is the census survey conducted in May/June 2000 covering all eligible households in the two regions selected for the program and the second, the baseline survey in August/September 2000, conducted for the randomized experiment prior to introduction of the subsidies. The data in the baseline survey includes detailed information on school enrolment, detailed direct and indirect costs (including fees, transport, books, uniforms, etc.) on schooling for those enrolled; health care utilization including consultations, type of provider, use of medication and hospitalization, direct and indirect costs of medical care and waiting times. However, the information on economic activity is sparse with only information on employment status, nature of employment, category of employment and hours worked. No information was collected on wages or income. All the above information was collected for all individuals of age 6 and over. Lack of information on income is substituted by detailed information collected on household expenditure and food consumption.

[3]The densities are estimated using the method of Li and Racine (2003) who use 'generalized product kernels' for mixed data. The bandwidths were set using the maximum likelihood cross validation

## 5.1 Variables

The census data provides information on the highest grade and level of education completed by all individuals aged 6 and over. The education of the household head can be mapped from this to the baseline survey. The census survey is also useful in trimming the sample to the program eligible children between the ages of 7-13 who have not completed grade 4. It also provides information on the distance to the nearest primary school.

School costs are only observed in the data for those children currently enrolled in school and must be estimated for all children in the sample. As mentioned above, both direct and indirect costs are observed and are aggregated into a single measure of costs. Human capital theory bases the family's choice of schooling on costs - both direct and indirect (opportunity costs), income and future returns to education (Becker, 1975). School costs are estimated using variables that capture these factors and include child characteristics - age of the child and gender. The lack of wage data poses a problem in estimating time costs of schooling. To overcome this, distance to school is used as a measure of opportunity cost of travel time. Family characteristic variables such as household expenditure, age, gender and years of schooling completed for the household head, number of children of school going age and number of adults in the family are included. Additionally, number of children under 5 is used as a measure of demand for child labour as often older children are expected to care for younger siblings. School costs can cause an endogeneity problem with household expenditure, to overcome this household expenditure net of school costs of the program eligible children is used as a measure of permanent income.

The estimation of the unobserved schooling outcomes under treatment $E(S_i|F_i = \tilde{F_j}, \theta_i = \tilde{\theta_j}, \lambda_i = \tilde{\lambda_j}, N_i = N_j, g_i = g_j, X_{hi} = X_{hj})$ is driven by the variables in the reduced form equations derived by the economic model ie. the variables determining the schooling decision are derived from $S^{**} = \Phi(\tilde{F}, \tilde{\theta}, \tilde{\lambda}, N, g; X_h)$ and include household expenditure (net of school costs and medical expenditure), a quadratic specification of age, number of children under 5, estimated school costs, medical expenses and years of education of the household head.

The baseline data covers 9747 individuals (both treatment and control) for 1581 households. This evaluation focuses on outcomes of children eligible for the schooling component of the program. Such households receive both the food transfer and the education transfer components of the program. The sample size for the purpose of this evaluation consists of 1786 children. Over half of this sample consists of families with more than one child eligible for the program.

# 6 Results

## 6.1 Estimating School Costs

Table 1: Estimating School Costs

| VARIABLES | (1) Probit-Boys Enrollment | (2) EEE-Boys School Costs | (3) Probit-Girls Enrollment | (4) EEE-Girls School Costs |
|---|---|---|---|---|
| age8 | 0.116 | 0.117 | 0.251 | 0.168 |
| | (0.151) | (0.139) | (0.160) | (0.103) |
| age9 | 0.265 | 0.180 | 0.544** | 0.385*** |
| | (0.161) | (0.147) | (0.168) | (0.103) |
| age10 | 0.174 | 0.0439 | 0.285 | 0.345* |
| | (0.164) | (0.124) | (0.172) | (0.172) |
| age11 | -0.00216 | 0.0458 | 0.202 | 0.470*** |
| | (0.164) | (0.133) | (0.179) | (0.116) |
| age12 | -0.00164 | 0.189 | 0.147 | 0.523** |
| | (0.172) | (0.137) | (0.188) | (0.197) |
| age13 | -0.554*** | 0.0256 | -0.159 | 0.205 |
| | (0.168) | (0.138) | (0.198) | (0.134) |
| HH Exp (adjusted) | 0.0000116** | 0.0000405*** | 0.0000122* | 0.0000264*** |
| | (0.00000447) | (0.00000345) | (0.00000593) | (0.00000359) |
| School dist | -0.00703*** | 0.00346** | -0.00806*** | 0.00230 |
| | (0.00178) | (0.00118) | (0.00173) | (0.00155) |
| No. of adults | -0.0336 | | 0.0536 | |
| | (0.0328) | | (0.0413) | |
| Children under5 | -0.172** | -0.224*** | -0.228*** | -0.0780 |
| | (0.0550) | (0.0398) | (0.0621) | (0.0430) |
| Children 7-13 | 0.0814 | -0.258*** | -0.0462 | -0.240*** |
| | (0.0463) | (0.0443) | (0.0557) | (0.0391) |
| HHH gender | 0.335 | | -0.0755 | |
| | (0.183) | | (0.198) | |
| HHH age | 0.00714 | | 0.00979 | |
| | (0.00529) | | (0.00562) | |
| HHH yrs of ed | 0.106** | | 0.180*** | |
| | (0.0381) | | (0.0436) | |
| HHH works | -0.0949 | | 0.186 | |
| | (0.173) | | (0.194) | |
| Constant | -0.0227 | -0.377* | -0.0752 | -0.288* |
| | (0.325) | (0.155) | (0.375) | (0.131) |
| $\lambda$ | | 0.289* | | 0.663** |
| | | (0.143) | | (0.204) |
| $\theta_1$ | | 1.242*** | | 1.564*** |
| | | (0.0887) | | (0.157) |
| $\theta_2$ | | 1.597*** | | 1.737*** |
| | | (0.106) | | (0.111) |
| Observations | 945 | 687 | 845 | 631 |

Robust standard errors in parentheses, clustered at the household level
*** p<0.01, ** p<0.05, * p<0.1

Table 1 shows the results from estimating the two part model for boys and girls. The probit participation model for both boys (1) and girls (3) show a similar pattern, with enrolment being most likely between the ages of 8 and 10 as compared to children aged 7 (reference category)and declining with older children. Boys

drop out earlier (above age 10) while girls aged 13 are less likely to enrol when compared to the reference group. This pattern follows most developing countries where many children enrol and stay in school only for a few years, dropping out between the ages of 11-13 to find employment. Household expenditure net of school costs (used as a proxy for income) and education of the household head are significant and have a positive impact on enrolment. As mentioned earlier the probit model includes the number of children under 5 years as a proxy for child labour. The estimates show similar negative magnitudes for boys and girls indicating having younger children in the household decreases the likelihood of enrolment. A similar effect of distance to the nearest school is observed, with children being less likely to enrol if schools are further away. Enrolment probabilities differ for boys and girls depending on the gender and the employment status of the head of the household. Girls are less likely to enrol if a male is head of the household, as is the case in 88% of the households in the sample. The direction of the coefficient for employment status is less intuitive as boys seem less likely to enrol if the household head is employed. This result is probably due to the nature of employment, with about 85% of the sample being involved in farm activities. The last two variables though not significant in the model do indicate the presence of a gender gap from additional opportunity costs for boys and cultural differences that contribute to the differences in schooling.

Columns (2) and (4) of Table 1 provide results from the second part of the two part model using the *extended estimating equations model (EEE)* (Basu and Rathouz, 2005) for school costs [4] . Boys in the reference category (age 7) face the highest school costs. At other ages there is no significant impact on school costs. For girls however, school costs increase with age. Families with greater wealth (household expenditure) tend to spend more on education, although more on the boys than the girls. In both cases children of the same age and children under five is significant (except for girls -children under5) and negative. This is intuitive in the sense that sharing of resources reduces the costs per child as the number of school age children increases.

In Column (2) for the boys sample the link parameter is estimated to be $\lambda = 0.289$ (95% C.I: 0.01, 0.57). The variance function represented by $\theta_1 = 1.2$ (95% C.I:1.07 ,1.42) and $\theta_2 = 1.6$ (95% C.I:1.39 , 1.80) is close to a gamma distribution. Column (4) provides the estimates for the sample of girls. In this case with $\lambda = 0.66$ (95% C.I: 0.26, 1.06), the link function is close to a square root link. The values $\theta_1 = 1.5$ (95% C.I:1.26 ,1.87) and $\theta_2 = 1.74$ (95% C.I:1.51, 1.95) again suggest a gamma distribution.

[4] An alternative approach to the EEE model would be to use a *generalized linear model* with a specified link function and distribution. However, failure to specify the correct link function results in misspecification of the model. To avoid such misspecifications, the EEE approach was used since it does not require an *a priori* assumption of a link function or distribution. This approach 'helps to identify an appropriate link function and to suggest an underlying distribution for a specific application but also serves as a robust estimator when no specific distribution for the outcome measure can be identified' Basu and Rathouz (2005).

## 6.2 Predicting Impacts

The empirical specification of the Klein and Spady model described in section 5.2 is used to predict unobserved school enrolment ($S^{**} = \Phi(\tilde{F}, \tilde{\theta}, N, g; X_h)$) under the treatment, accounting for the age of the child and a quadratic specification of age, number of children under 5, number of children between 7 and 14 years, number of adults, years of education of the head of the household, household expenditure, school costs, distance to the nearest primary and secondary school. Figures 4(a) and 4(c) illustrate the observed data from the comparison group ($S_j$ in equation 10) along with the Klein and Spady predictions for the extrapolated unobserved outcomes under treatment ($S_i$ in equation 10). Figures 4(b) and 4(d) compare predicted outcomes from the Klein and Spady model with those observed in the 2001 follow-up survey of the experiment. A comparison of Figures 4(b) and 4(d) shows that the extrapolated outcomes are quite close to the observed follow-up data for both boys and girls.

The estimator in equation 10 matches baseline program eligible children with characteristics $(\tilde{F}, \tilde{\theta}, X_h)$ with other baseline program eligible children with characteristics $(F, \theta, X_h)$. The estimated treatment effect is only valid for those families within the region of common support defined by equation 16. Figures 5(a), 5(b),5(c)and 5(d) compare the distributions of the variables included in the matching before and after trimming is implemented in the Klein and Spady estimator. They show that as required trimming eliminates observations where the density is very low, this translates to the ends of the right-tail of the distributions i.e families with very high household expenditure or school costs for whom matches are unlikely to be available are dropped from the estimation of treatment effects.

Table 2: Predicted Impact

| | (1) | (2) | (3) |
|---|---|---|---|
| VARIABLES | Predicted Impact | Sample sizes[@] | Experimental impact |
| Boys 7 -13 | 0.19*** | 859 / 876 | 0.19*** |
| | (0.0222) | | |
| Girls 7 -13 | 0.21*** | 754 / 767 | 0.20*** |
| | (0.0219) | | |
| Boys & Girls 7-9 | 0.17*** | 829 / 844 | 0.23*** |
| | (0.0255) | | |
| Boys & Girls 10-13 | 0.15*** | 786 / 799 | 0.15*** |
| | (0.0316) | | |

@ treatment observations after trimming, total number of observations.
Bootstrapped standard errors clustered at the comarca level (500 reps).
*** p<0.01, ** p<0.05, * p<0.1

The predicted impacts are listed in column (1) of Table 2 along with corresponding results from the *ex post* evaluation of RPS (column 3) [5]. This paper evaluates the impact of RPS using both treatment and control group data at the baseline as a single cross-section rather than just control group data as in the case of **?**. A comparison of the *ex ante* and *ex post* outcomes show that the *ex ante* approach predicts very closely the overall program impact for both boys and girls and is statistically significant, with one year of conditional cash transfers having a positive impact on enrolment of both boys and girls. The estimated impact for boys is 0.19 and accurately predicts the results of the ex-post evaluation. The one-year cash transfer increased enrolment of girls by 20 percentage points as compared to 21 percentage points from the experimental evaluation. In comparing the enrollment between boys and girls, girls continue to have higher enrolment rates even after 1 year of the program. At the baseline 75% of program eligible girls were already enrolled in school as compared to 72% of boys. The difference in opportunity costs could be a factor in explaining this difference as girls could be more likely to enrol possibly due to lower opportunity costs as compared to boys.

To examine further the predictions, the impacts are analysed by subgroups of age. The same model specification used for the boys and girls is used to estimate *ex ante* impacts for two sub-groups- children below 10 and those 10 years old and above. The ex-ante evaluation estimates an impact for children 10 years old and above as .15 which accurately predicts the experimental results. In the case of children below 10 years, the experimental evaluation shows a very large 23% rise in enrolment. The *ex ante* estimates for the same age group are also large and postive but are lower in magnitude at 17% when compared to the experimental estimates. In general the RPS program shows large and postive impacts across boys and girls and for different age groups. However, the impacts on children of younger ages is greater than for older children. The reduced form behavioural model approach applied here performs well in predicting one-year *ex ante* impacts of the RPS program.

# 7 Conclusion

This paper presents an *ex ante* evaluation of Nicaragua's CCT program Red de Protección Social. It applies the methods proposed by Todd and Wolpin (2010) on using reduced form estimation of behavioural models to carry out *ex ante* evaluations of social programs. The key requirement in this approach is that the preferences remain the same before and after the program so that the impact of the program is captured by

[5]The ex-post evaluation results from the published report of the evaluation of RPS provide only the overall impact (i.e row 1 of Table 2).The other values were calculated for this evaluation.

a change in the magnitude of the exogenous variables resulting from an introduction of the program. This paper extends their approach to jointly model both education and health outcomes and presents results from the schooling component of the program.

The model uses a health production framework and considers the influence of both direct and opportunity costs of schooling. Variation in the policy variable (school costs) and full income of the household is exploited to estimate program impacts on school enrolment. Empirically the model is implemented using a semi-parametric single index framework that allows for an increase in the dimensionality of the covariate vector. The outcome, school enrolment, is binary and the semi-parametric estimator proposed by Klein and Spady is used to predict the unobserved outcomes under treatment. The data set combines baseline data from the RPS experiment along with some information from the census survey . The baseline data is used as a single cross-section combining both control and treatment groups. Comparing the predicted estimates with the experimental outcomes shows that the predictions all have the same direction as the experimental impact. The predictions for overall impact of the program for boys, girls and the age group - 10 years and above are very close in magnitude to the experimental impact. The prediction for younger children however is lower than the experimental impact but still shows a large positive effect of a one year cash transfer. The empirical approach used relies on selection on observables performs well when the observables are fully captured. In general, in keeping with the findings from the experiment, the *ex ante* evaluation finds a significant and large overall impact of RPS on the target population.

# References


Attanasio O, Meghir C, Santiago A (2005) Education choices in mexico: using a structural model and a randomized experiment to evaluate progresa. Open access publications from university college london, University College London, URL http://ideas.repec.org/p/ner/ucllon/http–eprints.ucl.ac.uk-14750-.html

Basu A, Rathouz PJ (2005) Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostat 6(1):93–109

Becker GS (1965) A theory of the allocation of time. The Economic Journal 75(299):493–517, URL http://www.jstor.org/stable/2228949

Becker GS (1975) Human Capital. New York: National Bureau of Economic Research

Chamberlain G (1986) Asymptotic efficiency in semi-parametric models with censoring. Journal of Econometrics 32(2):189–218, URL http://ideas.repec.org/a/eee/econom/v32y1986i2p189-218.html

Cosslett SR (1987) Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. Econometrica 55(3):559–585

Fernández AI, Rodríguez-Poo JM (1997) Estimation and specification testing in female labour participation models: Parametric and semiparametric methods. Econometric Reviews 16(2):229–247

Gerfin M (1996) Parametric and semi-parametric estimation of the binary response model of labour market participation. Journal of Applied Econometrics 11(3):321–339

Hayfield T, Racine JS (2008) Nonparametric econometrics: The np package. Journal of Statistical Software 27(5), URL http://www.jstatsoft.org/v27/i05/

Heckman JJ, Ichimura H, Todd PE (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies 64(4):605–54

Horowitz JL (1993) Semiparametric estimation of a work-trip mode choice model. Journal of Econometrics 58(1-2):49–70

Horowitz JL (1998) Semiparametric methods in econometrics. New York: Springer-Verlag

Ichimura H (1993) Semiparametric least squares (sls) and weighted sls estimation of single-index models. Journal of Econometrics 58:71–120

Ichimura H, Taber CR (2000) Direct estimation of policy impacts. NBER Technical Working Papers 0254, National Bureau of Economic Research, Inc, URL http://ideas.repec.org/p/nbr/nberte/0254.html

IFPRI (2001a) Evaluation design for the pilot phase of the nicaraguan red de protección social. report submitted to the red de protección social, International Food Policy Research Institute

IFPRI (2005) Nicaragua:red de protección social (rps) evaluation dataset, 2000-2002. Tech. rep., Washington,D.C.: International Food Policy Research Institute (IFPRI) (datasets). http://www.ifpri.org/dataset/nicaragua

Klein RW, Spady RH (1993) An efficient semiparametric estimator for binary response models. Econometrica 61(2):387–421

Li Q, Racine J (2003) Nonparametric estimation of distributions with categorical and continuous data. Journal of Multivariate Analysis 86(2):266–292

Maluccio JA, Flores R (2005) Impact evaluation of a conditional cash transfer program: the nicaraguan red de proteccin social. Research reports 141, International Food Policy Research Institute (IFPRI), URL http://ideas.repec.org/p/fpr/resrep/141.html

Manning WG (1998) The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 17(3):283 – 295

Mullahy J (1998) Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journal of Health Economics 17(3):247 – 281

Todd PE, Wolpin KI (2006) Assessing the impact of a school subsidy program in mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. American Economic Review 96(5):1384–1417, URL http://ideas.repec.org/a/aea/aecrev/v96y2006i5p1384-1417.html

Todd PE, Wolpin KI (2010) Ex ante evaluation of social programs. Annales d'Economie et de Statistiques in press

White H (2006) Impact evaluation: The experience of the independent evaluation group of the world bank. Tech. Rep. Report No. 1111, Washington D.C, The World Bank

World B (1998) Nicaragua poverty assessment: Challenges and opportunities for poverty reduction. Tech. Rep. Report No. 20488-NI. Washington, D.C:, The World Bank

| PROGRAM REQUIREMENT | HOUSEHOLD TYPE | | | |
|---|---|---|---|---|
| | Households with no targeted children (A) | Households with children aged 0-5 (B) | Households with children aged 7-13 who have not completed 4th grade (C) | (B) + (C) |
| Attend bimonthly health education workshops | ✓ | ✓ | ✓ | ✓ |
| Bring children to prescheduled healthcare appointments | | | | |
| Monthly (0-2 years) | | ✓ | | ✓ |
| Bimonthly (2-5 years) | | | | |
| Adequate weight gain for children under 5[a] | | ✓ | | ✓ |
| Enrollment in grades 1 to 4 of all targeted children in the household | | | ✓ | ✓ |
| Regular attendance (85 percent, i.e., no more than 5 absences every two months without valid excuse) of all targeted children in the household | | | ✓ | ✓ |
| Promotion at end of school year[b] | | | ✓ | ✓ |
| Deliver teacher transfer to teacher | | | ✓ | ✓ |
| Up-to-date vaccination for all children under 5 years[b] | | ✓ | | ✓ |

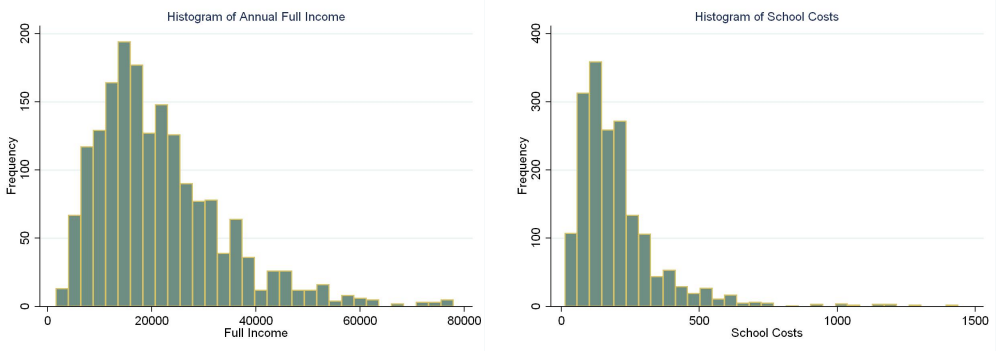
a. The adequate weight gain requirement was discontinued in Phase 11, starting in 2003
b. Condition was not enforced.

Figure 1: RPS Eligibility and Requirements. Source: Maluccio and Flores 2005

**Nicaraguan RPS eligibility and benefits in Phase 1**

| | Program components | |
|---|---|---|
| | **Food security, health and nutrition** | **Education** |
| Eligibility<br>Geographic targeting | All households | All households with children aged 7-13 who have not completed fourth grade of primary school |
| Demand-side benefits<br>Monetary transfers | *Bono alimentario*<br>(food security transfer)<br>C$2,880 per household per year<br>(US$224) | *Bono escolar*<br>(school attendance transfer)<br>C$1,440 per household per year<br>(US$112)<br><br>*Mochila escolar*<br>(school supplies transfer)<br>C$275 per child beginning of school year<br>(US$21) |
| Supply-side benefits<br>Services provided and monetary transfers | Health education workshops every 2 months<br>Child growth and monitoring<br>Monthly: Newborn to 2-year-olds<br>Every 2 months: 2- to 5-year-olds<br>Provision of antiparasite medicine, vitamins, and iron supplements<br>Vaccinations (newborn to 5-year-olds) | *Bono a la oferta*<br>(teacher transfer)<br>C$80 per child per year given to teacher/school (US$6) |

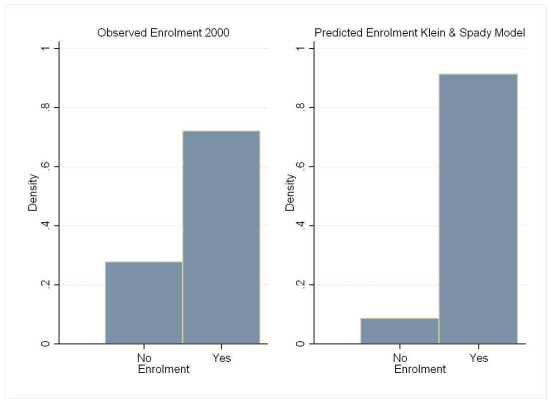
Figure 2: RPS Transfers. Source: Maluccio and Flores 2005



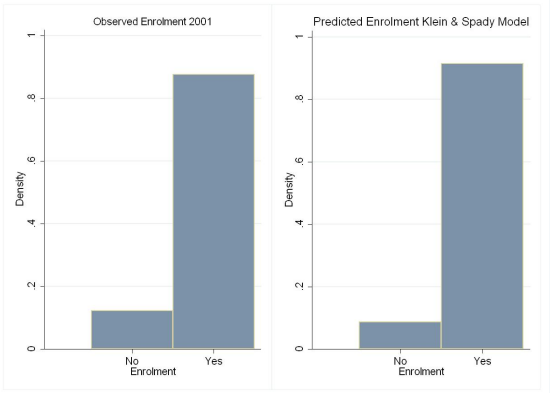

(a) Household Expenditure

(b) School Costs

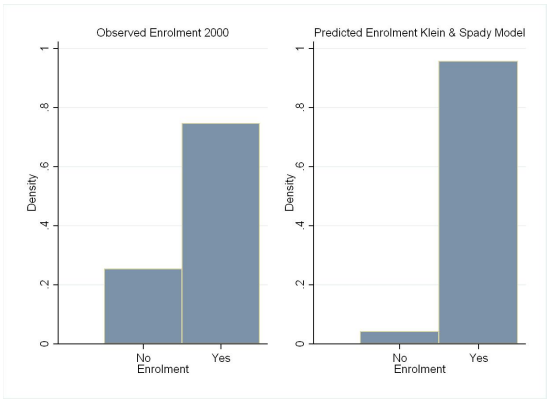
Figure 3: Data Variation



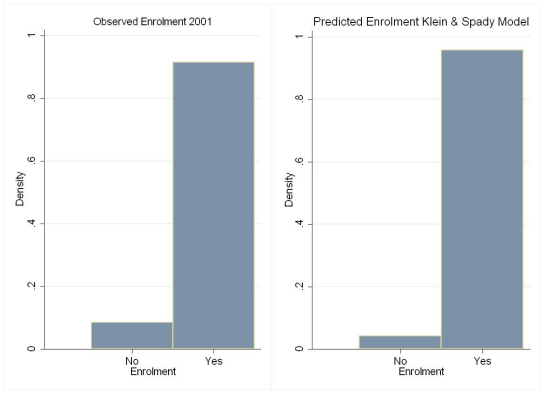

(a) Observed Baseline Outcomes with Predicted Outcomes for Sample of Boys




(b) Observed Follow up Outcomes with Predicted Outcomes for Sample of Boys

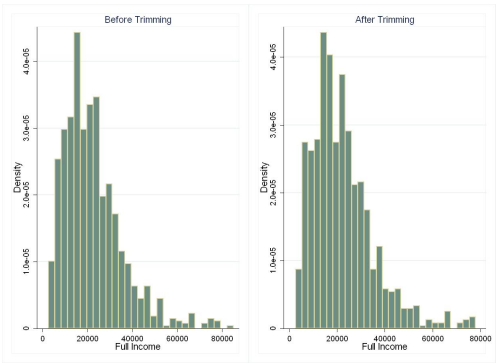



(c) Observed Outcomes and Predicted Outcomes for Sample of Girls

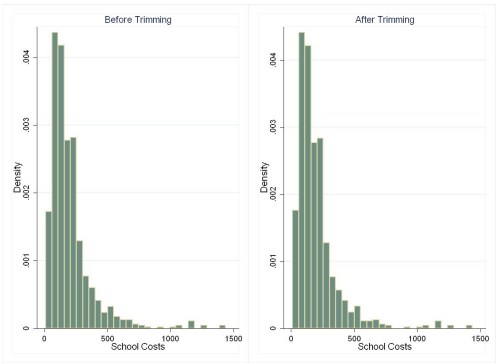



(d) Observed Follow up Outcomes with Predicted Outcomes for Sample of Boys

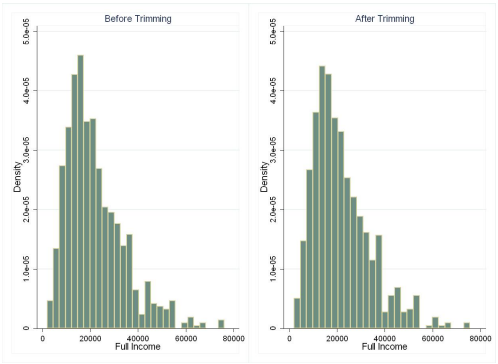
Figure 4: Comparing Observed and Predicted Outcomes



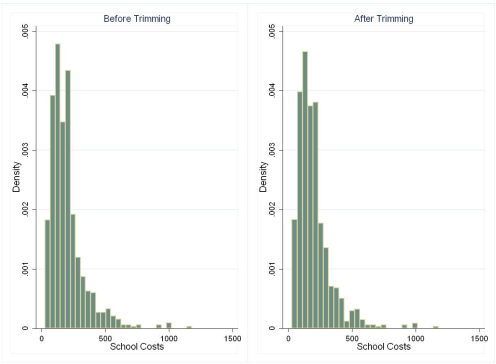

(a) Full Income for Sample of Boys




(b) School Costs for Sample of Boys




(c) Full Income for Sample of Girls




(d) School Costs for Sample of Girls

Figure 5: Trimming Klein and Spady estimations