1. Critical assessment.

Summary

This paper estimates the average and marginal returns to schooling. The research is motivated by the impressive record of educational expansion in Indonesia since 1970s and by the importance of understanding the impact of education on earnings of those affected by these expansions for policy development.

A standard potential outcome framework is applied to modeling schooling. For simplicity of the model, schooling (*S*) is presented as a dummy variable, which equals 1 if individual completed upper secondary school or higher and 0 if individual completed only lower secondary school or below. Given two levels of schooling, there are two potential outcomes Y_1 and Y_0, respectively, which result in the following model:

$$\begin{aligned} Y_1 &= \alpha_1 + X\beta_1 + U_1 \\ Y_0 &= \alpha_0 + X\beta_0 + U_0 \end{aligned} \qquad (1)$$

$$S = 1 \text{ if } Z\gamma - U_s > 0 \qquad (2)$$

In theory individuals consider gains and costs from schooling, when they decide whether to enroll to upper secondary school or not (2).

Empirical analysis is performed using semiparametric selection model. The identification strategy is given by geographic variation of access to secondary school. This variation as well as a rich set of control variables is used to predict propensity score, which is used for local IV method implementation. To characterize the heterogeneity of returns, the marginal treatment effect (MTE) is estimated, which allows to measure the returns to schooling for individuals with different levels of observables and unobservables. In addition, such parameters of interest as average treatment effect (ATE), average treatment effect for treated (ATT) and for not treated (ATU) are constructed as weighted average of MTE.

The main finding of this paper is that the return to upper secondary schooling for the marginal person is significant, but much lower than the returns for the average person. (14.2% vs. 26.9% per year of schooling).

Important assumptions and modeling decisions

The advantage of using local IV method is that no distribution assumptions on unobservables should be imposed. The only necessary assumption is independence of X (control variables) and Z (instrumental variables) from error terms U_1, U_0 and Us. It implies that the shape of the MTE curve does not vary with X and the MTE can thus be identified over the unconditional support of the propensity score. This is a strong assumption to make.

To identify the parameters of interest, the IV assumptions such as validity, excludability and random assignment should be satisfied. Propensity score is used as the instrument. It is estimated by the logit regression of binary schooling variable on the distance to upper secondary school (instrumental variable) and controls X. Such instrument is a strong predictor of the enrollment to secondary school, that is why the validity assumption is easily satisfied, while randomness and excludability are less clear and will be discussed further.

Limitations of the data

The sample consists of 2608 working age (25-60 years old) males from 321 randomly selected villages spread among 13 Indonesian provinces. Females are excluded from the sample because of low labor force participation. The results of this analysis, therefore, do not provide any information regarding the returns to schooling for women. In addition, almost 90% of the population is Muslim, which makes the external validity of the estimates even more unclear. Another issue is type of the data used, which is a survey. The dependent variable of this analysis, log hourly wages, is constructed from self-reported monthly wages and hours worked per week. People tend to give inaccurate information regarding their wages in interviews, that is why data may contain some measurement error. The measurement error in the dependent variable, however, is not so problematic and leads to less precise, but still unbiased and consistent estimates. Measurement error bias could also appear because of measurement error in schooling variable. This would be a problem if school participation is intermittent may introduce some recall bias.

This sample, though, contains rich data on many variables, which is important for the flexibility of the model.

Weaknesses and strong points of the econometric analysis

The strong feature of this empirical analysis is flexibility of the model estimation, which is important in case of heterogeneity of effects. In addition, the MTE estimation is much more informative way of exploiting a continuous variable instrument which, unlike IV and control function estimations, allows to identify a variety of parameters, e.g. ATE, ATT, ATU and PRTE, T. Cornelissen et al. (2015). This paper also suggests the innovative approach of the weights estimation used to construct different parameters of interest. Simulating the weights using the estimated parameters resolves the problem of the need to estimate the multidimensional conditional density function.

The potential weakness of this analysis lies in its instrument. An instrumental variable used in the estimation is current distance to school. To identify the parameters of interest, the IV assumptions should be satisfied. The assumption of validity is satisfied. Excludability assumption, though, is more difficult to satisfy and it could not be tested. There are two problems with the chosen instrumental variable. Firstly, families and schools may not randomly locate in Indonesia, which will lead to the violation of random assignment of the instrument. Secondly, there is possible reverse causality in the first stage regression. Namely, educated individuals may choose to move to urban areas, that also have more schools. These issues are solved by including rich set of control variables, conditioning on which, excludability of the instrumental variable is assumed to be satisfied. Such procedure creates, however, a limitation. As propensity score could never be observed, but still should be estimated, it will have an estimation error, which should be taken into account when estimating standard errors.

2. Replication of Tables 1 to 4.

TABLE 1

For the construction of the first table, one would need to create dummy variables for religion, father's education mother's education and province. These dummy variables will be used to calculate means of variables for the treatment and control groups.

```
tab religion, gen(religionx)
tab feduc, gen(feducx)
tab meduc, gen(meducx)
tab province, gen(provincex)
```

To simplify construction of the tables and to make results more readable, all variables are labeled:

```
label define religion 0 "Muslim", modify

label var religionx1 "Muslim"
label var religionx2 "Protestant"
label var religionx3 "Catholic"
label var religionx4 "Other"

label var feducx1 "Father: uneducated"
label var feducx2 "Father: elementary"
label var feducx3 "Father: secondary"
label var feducx4 "Father: missing"

label var meducx1 "Mother: uneducated"
label var meducx2 "Mother: elementary"
label var meducx3 "Mother: secondary"
label var meducx4 "Mother: missing"

label var provincex1 "North Sumatra"
label var provincex2 "West Sumatra"
label var provincex3 "South Sumatra"
label var provincex4 "Lampung"
label var provincex5 "Jakarta"
label var provincex6 "West Java"
label var provincex7 "Central Java"
label var provincex8 "Yogyakarta"
label var provincex9 "East Java"
label var provincex10 "Bali"
label var provincex11 "West Nussa Tengara"
label var provincex12 "South Kalimanthan"
label var provincex13 "South Sulawesi"
```

Then sample statistics are calculated using command “summarize” and results are stored using “eststo”:

```
eststo upper: quietly estpost summarize ///
     learnhr00 kmsmp kmsd age religionx* feducx* meducx* ///
rural provincex* if dschool==1
eststo less: quietly estpost summarize ///
     learnhr00 kmsmp kmsd age religionx* feducx* ///
meducx* rural provincex* if dschool==0
```

Finally, the table is replicated in the following way:

```
esttab upper less, ///
cell(mean(fmt(%9.3f))) varwidth(20) label ///
nonumbers modelwidth(30 30) collabels(none) ///
title("Table1:Sample statistics for treated and control groups") ///
mtitles("Upper secondary or higher" "Less than upper secondary")
```

Not all the obtained sample statistics are the same as in the paper. There are some small discrepancies in the instrumental variable – distance to school (km), in father’s and mother’s education for uneducated category in the treatment group as well as in the control group.

TABLE 2

For the logit regression, one more variable should be added. It is a square of the age, which is generated in the following way:

```
gen age2=age^2
```

The first column of Table 2 shows the estimates of the logit regression of upper secondary school attendance on control variables.

```
qui logit dschool kmsmp age age2 i.religion i.feduc ///
i.meduc rural kmsd i.province
estimates store logit
```

Test for joint significance of instrument:

```
test kmsmp
```

Mean of dependent variable:

```
sum dschool
di "Mean of dependent variable=" r(mean)
```

The second column of Table 2 presents average derivatives (calculated at means of control variables). In this case, though, logistic regression includes instrumental variable and its interactions with age, religion, parents' education and rural residence to make a model more flexible.

```
logit dschool kmsmp c.kmsmp#c.age c.kmsmp#i.religion ///
c.kmsmp#i.feduc ///
c.kmsmp#i.meduc c.kmsmp#c.rural ///
age age2 i.religion i.feduc i.meduc c.rural kmsd ///
i.province, cluster(commid00)

margins, dydx(*) at((mean)) post
estimates store margins
```

Finally, Table 2 is constructed in the following way:

```
esttab logit margins, ///
b(%9.3f) se(%9.3f) star(* 0.10 ** 0.05 *** 0.01) ///
drop(*province *0.religion *0.feduc *3.feduc *0.meduc *3.meduc) ///
title("Table 2: Upper school decision model") ///
mtitles("Coeff" "Average derivatives") varwidth(30) nonumber label
noobs nogaps
```

The coefficients in the table again are not absolutely the same as in the paper. One reason could be the discrepancies in the instrumental variable.

TABLE 3

In the first column of Table 3, coefficients of the OLS estimation is presented, which are estimated in the following way:

```
reg learnhr00 dschool age age2 i.religion i.feduc ///
```

```
i.meduc rural kmsd i.province, cluster(commid00)
estimates store OLS
```

The annualized version of the parameter is obtained in the following way:

```
di "Upper secondary (annualized) OLS=" _b[dschool]/7.79
```

The difference in the average years of schooling 7.79 is taken from the paper.

The second column presents coefficients of the 2sls IV estimation. The instruments used in the estimation are distance to secondary school and interactions with parental education, religion and age. Followed by the calculation of the annualized version of the parameter of interest.

```
ivregress 2sls learnhr00  (dschool=c.kmsmp ///
c.age#c.kmsmp i.religion#c.kmsmp i.feduc#c.kmsmp ///
i.meduc#c.kmsmp) ///
age age2 i.religion i.feduc i.meduc rural kmsd ///
i.province, cluster(commid00)
estimates store IV

di "Upper secondary (annualized) IV=" _b[dschool]/7.79
```

The final table is constructed in the following way.

```
esttab OLS IV, ///
b(%9.3f) se(%9.3f) star(* 0.10 ** 0.05 *** 0.01)  r2 ///
drop(*province *0.religion *0.feduc *3.feduc *0.meduc *3.meduc) ///
title("Table 3: Annualized OLS and IV estimates") ///
mtitles("OLS" "IV") varwidth(30) nonumber label nogaps
```

TABLE 4

This table presents nonparametric and normal selection estimates of chosen parameters. It is possible to calculate marginal treatment effects (MTE) using both normal selection model and semiparametric model with a help of Stata package *margte,* S.Brave et al.(2014).

(1) Semiparametric model (double residual method of Robinson)

```
margte learnhr00 age age2 religion feduc meduc ///
rural kmsd province, ///
treatment(dschool  kmsmp age age2 i.religion i.feduc ///
i.meduc rural kmsd i.province) ///
semiparametric kernel(gaussian) xbwidth(0.27) link(logit)
```

(2) Normal selection model (switching regression parametric model):

```
margte learnhr00 age age2 religion feduc meduc ///
rural kmsd province, ///
treatment(dschool  kmsmp age age2 i.religion ///
i.feduc i.meduc rural kmsd i.province)
```

To calculate all other parameters such as ATT and ATN, grid of values of unobservable term *V* should be created. In this paper, new methodology is used. Due to multiple control variables, estimation of parameters requires estimation of conditional densities, where the conditioning set is of high dimensionality. That is why instead a simulation method is used. Unfortunately, I was not able to reproduce the weights and calculate parameters.

3. Extension of the analysis.

Matching could be an alternative method to estimate returns to schooling***.*** Matching method requires an extensive set of observable characteristics, on which to match. Available dataset provides possibilities to implement matching. It is a non-parametric approach of identifying the treatment impact on outcomes. Matching assumes that one can proceed as in a RCT, once controlled for all relevant characteristics. The central issue in the matching method is choosing the appropriate matching variables.

To estimate parameters of interest, two following assumptions should be satisfied:

(1) Conditional Independence assumption

It assumes that all the outcome-relevant differences between treated and untreated are captured in their observable attributes.

In case of applying matching to the question estimated in the paper Carneiro et al. (2017), all predetermined confounding variables such as age, religion, parental education, distance to the health post, indication variable for living in the village at age 12 could be controlled for. One could believe that CIA is a reasonable assumption in this case, as there is a relatively rich array of variables on which to match. Rich dataset, though, does not guarantee that conditional independence assumption holds. Motivation and ability, which affect potential wage, are not observed and may correlate with schooling choice. In case, individuals with high motivation, would choose to enroll to upper secondary school, estimates will be upward biased. If individuals with lower secondary school education or lower are more motivated to work and earn money, estimates will be downward biased. In such cases, conditional independence assumption will be violated.

(2) Common support.

There should be some untreated individuals. This could be tested.

Before matching control and treatment groups, one can look at t-tests, which check the equality of means.

```
global x age age2 religion feduc meduc rural kmsd province
foreach v of var $x {
qui ttest `v', by(dschool)
di "`v'" _col(15) %10.0g r(mu_1) %10.0g r(mu_2 ) %6.3f r(p)
}
```

By looking on means of treated and control groups, it could be easily seen that they are quite equal (Table 1).

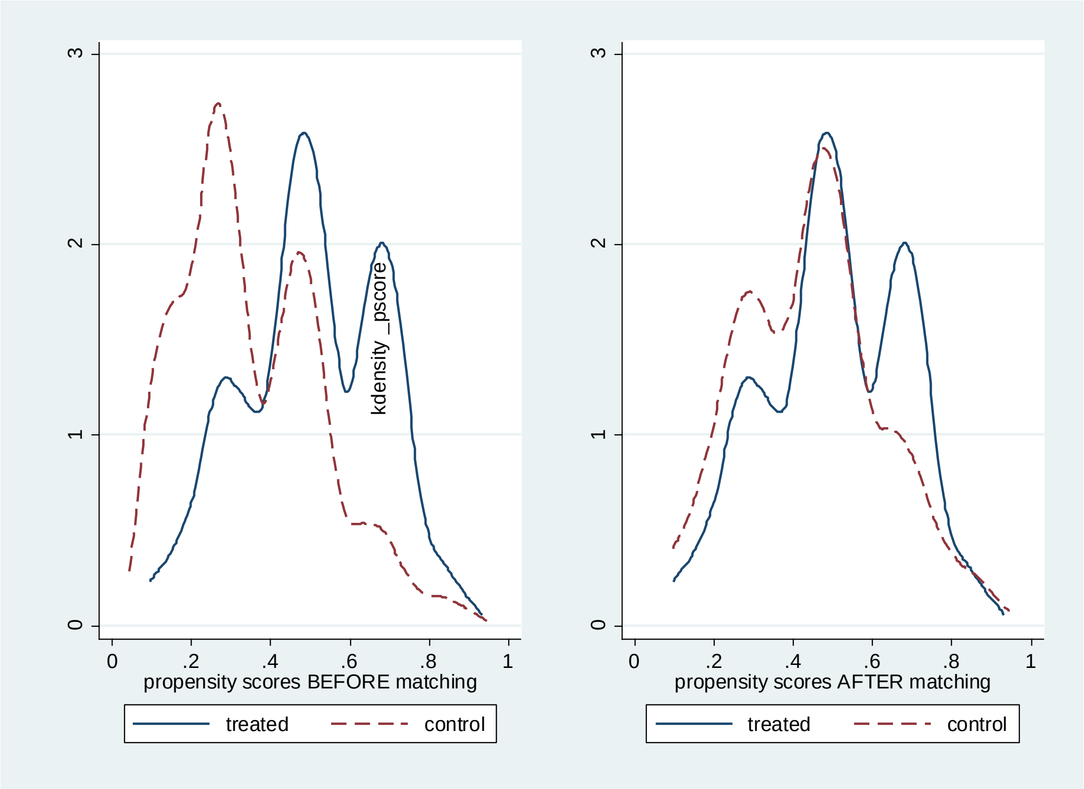
Table 1. T-test of equality of the means between treated and control groups

Variables	Treated	Untreated	p-value
Age	38.67	37.06	0.000
Age (squared)	1585.66	1451.08	0.000
Religion	0.17	.29	0.000
Father’s education	0.74	1.22	0.000
Mother’s education	0.85	1.18	0.000
Rural	0.48	0.24	0.000
Distance to health post	1.08	0.89	0.000
Province	33.22	34.16	0.077

It is common to use propensity score to match individuals in treated and control groups, which allows to overcome dimensionality issue. The parameters of interest could be estimated using 1 nearest neighbor propensity score matching, which could be implemented by the following Stata code.

```
psmatch2 dschool $x, out(learnhr00)
//Assess matching quality
pstest $x
```

One can also check whether propensity scores were matched in appropriate way by looking at the graphs below, which illustrate propensity scores distributions before and after matching.



```
twoway (kdensity _pscore if _treated==1) ///
(kdensity _pscore if _treated==0, lpattern(dash)), ///
legend( label( 1 "treated") label( 2 "control" ) ) ///
xtitle("propensity scores BEFORE matching") saving(before)

gen match=_n1
replace match=_id if match==.
duplicates tag match, gen(dup)
twoway (kdensity _pscore if _treated==1) ///
(kdensity _pscore if _treated==0 ///
& dup>0, lpattern(dash)), legend( label( 1 "treated") ///
 label( 2 "control" )) ///
xtitle("propensity scores AFTER matching") saving(after)

graph combine before.gph after.gph, ycommon
```

Now parameters of interest could be estimated

```
psmatch2 dschool age age2 i.religion i.feduc i.meduc ///
rural kmsd i.province, out(learnhr00) ate logit
```

Annualized estimates are presented in Table 2.

Table 2. Estimates of average returns to upper secondary schooling

Parameter	Nonparametric estimate	Normal selection model estimate	Matching estimate
ATT	0.218	0.203	0.08
ATE	0.138	0.067	0.10
ATU	0.081	-0.029	-0.11

Note: Estimates of nonparametric regression and normal selection model were taken from Carneiro et al. (2017)

The results of matching are very different from the estimates of semiparametric and normal estimation model. They are much lower. One should also check other matching methods. In addition, it is not clear whether the conditional independence assumption will be satisfied in this case.

Advantage of matching is that it is a non-parametric method. No assumptions are needed on the form of the outcome equation, decision process or unobservable term. Therefore, this method is flexible and intuitive. There are disadvantages, though. Matching requires much data. In addition, the estimated effect could be redefined as the mean treatment effect for those falling within the common support, which could change parameter being estimated. Lastly, it is not asymptotically efficient.

The other way to extent the analysis is to look at sub-populations, e.g. different age groups. The sample could be divided into two groups: individuals with age equal to or below 37 (which is a median age in the sample) and individuals above 37. The descriptive statistics are presented in Table 3. As it was mentioned in the paper, the age could be a proxy of experience. One may notice that average wages for in, individuals older than 37 are higher for both treated and control groups. Unfortunately, the number of observation in each group is not very high and, therefore, even if one runs semiparametric model to estimate treatment effects, the results will not be robust and could not be reliable.

Table 3: Sample statistics for the treatment and control groups by age group

	Upper (<=37 years old)	**Upper (>37 years old)**	**Less (>=37 years old)**	**Less (<37 years old)**
Log of hourly wages	7.923	8.568	7.406	7.551
Distance to school (km)	0.967	0.921	1.500	1.320
Distance to health post (km)	0.902	0.870	1.118	1.042
Age	30.555	45.795	30.500	46.401
Muslim	0.868	0.849	0.947	0.908
Protestant	0.043	0.058	0.018	0.026
Catholic	0.021	0.039	0.005	0.011
Other	0.068	0.054	0.030	0.055
Father: uneducated	0.096	0.214	0.304	0.481
Father: elementary	0.481	0.533	0.565	0.452
Father: secondary	0.394	0.244	0.070	0.052
Father: missing	0.029	0.009	0.061	0.014
Mother: uneducated	0.133	0.324	0.318	0.554
Mother: elementary	0.490	0.475	0.473	0.344
Mother: secondary	0.265	0.121	0.026	0.018
Mother: missing	0.111	0.080	0.184	0.084
Rural household	0.236	0.244	0.531	0.424
North Sumatra	0.061	0.052	0.055	0.070
West Sumatra	0.039	0.058	0.046	0.069
South Sumatra	0.048	0.048	0.024	0.038
Lampung	0.014	0.017	0.032	0.022
Jakarta	0.201	0.153	0.091	0.098
West Java	0.162	0.188	0.214	0.163
Central Java	0.084	0.086	0.168	0.160
Yogyakarta	0.082	0.106	0.057	0.051
East Java	0.119	0.123	0.173	0.186
Bali	0.059	0.052	0.023	0.052
West Nussa Tengara	0.051	0.048	0.058	0.038
South Kalimanthan	0.040	0.039	0.019	0.022
South Sulawesi	0.039	0.030	0.041	0.029
Observations	622	463	740	783

This is how this table could be replicated in Stata.

```
eststo upper_younger: quietly estpost summarize ///
     learnhr00 kmsmp kmsd age religionx* feducx* meducx* rural ///
provincex* if dschool==1 & age<=37
eststo upper_older: quietly estpost summarize ///
     learnhr00 kmsmp kmsd age religionx* feducx* meducx* rural ///
```

```
provincex* if dschool==1 & age>37
eststo less_younger: quietly estpost summarize ///
     learnhr00 kmsmp kmsd age religionx* feducx* meducx* rural ///
provincex* if dschool==0 & age<=37
eststo less_older: quietly estpost summarize ///
     learnhr00 kmsmp kmsd age religionx* feducx* meducx* rural ///
provincex* if dschool==0 & age>37

esttab upper_younger upper_older less_younger less_older ///
using "M:\byage7.rtf", ///
cell(mean(fmt(%9.3f))) varwidth(30) label nonumbers ///
modelwidth(10 10) collabels(none) ///
title("Table 1: Sample statistics for the treated and control") ///
mtitles("Upper (<=37 years old)" ///
"Upper(>37 years old)" ///
"Less(>=37 years old)" ///
"Less (<37 years old)")
```

Literature review

1. Cornelissen, T., Dustmann, C., Raute, A., Schönberg, U. (2015) “From LATE to MTE: alternative methods for the evaluation of policy interventions.” JEL: C26, I26

https://annaraute.files.wordpress.com/2013/09/late_to_mte-paper_june2016.pdf

2. Carneiro, P., Lokshin, M., and Umapathi, N. (2017) Average and Marginal Returns to Upper Secondary Schooling in Indonesia. J. Appl. Econ., 32: 16–36.

http://www.ucl.ac.uk/~uctppca/mte_nov27.pdf

3. Brave S., Walstrum, T. “Estimating marginal treatment effects using parametric and semiparametric methods” The Stata Journal (2014), Number 1, pp. 191–217

http://www.stata-journal.com/sjpdf.html?articlenum=st0331