NONPARAMETRIC BOUNDS ON RETURNS TO SCHOOLING IN SOUTH AFRICA

MARTINE MARIOTTI AND JÜRGEN MEINECKE

We nonparametrically estimate upper bounds on the average treatment effect of one additional year of schooling in South Africa for 1993 and 1998. The study uses the KwaZulu-Natal Income Dynamics Study (KIDS) panel data set. Compared to the existing parametric literature our upper bound is informative: the average treatment effect is bounded above by 7.1 percent and 8.7 percent for 1993 and 1998 and the standard errors are tight. Our results suggest that many parametric estimates are severely upwards biased, which results from unobserved heterogeneity.

Keywords: returns to schooling, partial identification, nonparametric estimation.

1. INTRODUCTION

One of the policies of Apartheid in South Africa (1950-1994) was the unequal distribution of education across race groups such that whites have historically attained high levels of education while Africans have attained low levels. An important consequence of the educational distribution was unequal opportunities for employment in skilled occupations, with Africans forced to work in lower skilled occupations because of their lower educational attainment (Mariotti 2009). As a result, a post Apartheid adjustment to education policy was called for in order to level the playing field with the expectation that higher education leads to higher incomes through employment in more skilled occupations.

Using nonparametric techniques, we estimate upper bounds on the average treatment effect of one additional year of schooling in South Africa. Our analysis consists of both a cross-sectional data set covering all geographical regions in the country as well as a panel data set focused on only one of South Africa’s nine provinces – KwaZulu–Natal. The cross-sectional data provides a snapshot of education and incomes in 1995 and 2000 while the panel data set allows us to determine the evolution of the return to education from 1993 to 1998. Comparing employed African males with eight years of education to those with zero years of education, we estimate an upper bound annual return of 7.1 percent in 1993 which increases to 8.7 percent in 1998 for KwaZulu–Natal, while across the entire country we find an upper bound of 8.7 percent in 1995 which decreases to 6.95 percent by 2000. These numbers are substantially lower than most parametric estimates.

A number of studies have attempted parametric estimations of the returns to education in post–Apartheid South Africa. A summary of these results shows a wide distribution of returns with some returns as high as 100 percent for secondary and tertiary education. The high returns are surprising when one considers an ad–hoc observation of educational attainment in South Africa. Such extremely
high returns might be expected to lead to a higher demand for education than currently observed. Furthermore, such high returns suggest that an obvious avenue for policy is to dramatically increase students’ access to secondary and tertiary education, an outcome we do not see.

One possible explanation for the discrepancy between what the parametric returns suggest and the actual demand for education is that these studies do not account for unobserved heterogeneity. We circumvent this problem by using Manski and Pepper’s (2000) nonparametric estimator which applies two very mild assumptions and does not require any conditional independence assumption regarding unobserved heterogeneity. One drawback of our approach is that we can only estimate an upper bound on the average treatment effect of education. However, with a sufficiently low estimate these bounds are still meaningful.

Only two parametric studies find returns lower than ours but we have reason to believe that our upper bound is a conservative estimate. We anticipate through the use of slightly more restrictive assumptions following the approach of Blundell et al. (2007), that we will able to further tighten the bounds.

2. LITERATURE REVIEW

One of the secondary benefits of the fall of Apartheid in 1994 has been an improvement in data collection that has provided researchers the opportunity to document the transformation of individual social and economic characteristics. In particular, a large body of work has documented changes in the return to education since the fall of Apartheid. The results are consistent in the racial hierarchy of returns, in that, firstly, Africans persistently earn a higher return for higher levels of education and, secondly, that higher levels of education earn a higher return across all race groups relative to no education. However, there is a large variation in the quantification of the returns with several studies finding surprisingly high returns.

Thomas (1996) provides a brief, useful account of the state of education in South Africa by 1991. Dividing the 1991 population census into cohorts he shows that whites attained higher levels of education than non-whites, and that levels of education have been increasing over time for all race groups. Indians and Coloureds have the second highest levels of educational attainment, with Africans acquiring the lowest amount of education. Following the Soweto school riots of 1976, the government increased expenditure on African education within South Africa (but not the homelands which were supposed to be funding their own students). Despite the increased expenditure, by 1991 African education continued to lag behind that of the other races.

Mwabu and Schultz (1996) use the Project for Standards of Living Survey Data (PSLSD) set of 1993 to measure the returns to education. Using the working age population, an OLS regression finds that the return to education is 16% for secondary education and 27% for higher education for Africans while for whites the comparative returns are 8% and 15%. The quantile regression results find that

\[ \text{The racial hierarchy arises most likely because the proportion of Africans with high levels of education—a legacy of the apartheid education system—is extremely low. This result is despite the restricted employment opportunities for highly educated Africans.} \]
the return to primary education for Africans is between 10% and zero, the return to some secondary education is between 10% and 18% (although the difference is not significant) and the return to higher education ranges from around 23% to around 30%. For whites, the return to primary education is zero, the return to secondary ranges from zero to 20% and the return to higher education ranges from around 7% to around 15%. They also apply the quantile regression approach to determine the direction of correlation between education and ability. The negative correlation between the two suggests that people substitute education for ability.

In a later paper, Mwabu and Schultz (2000) look again at returns to education and find that the returns are greater at higher levels of educational attainment for Africans than for whites. This is most likely a result of both the low quantity of Africans who have attained higher education as well as job reservation which means that Africans and whites do not compete for the same jobs and therefore do not compete for wages. Using the PSLSD, they spline education into three groups (primary, secondary and higher), apply OLS and find that the return to African education (for men) is 8.4% at the primary level, 15.8% at the secondary level and 29.4% at the tertiary level. For white men they find 0% for primary, 8.4% for secondary and 15.1% for tertiary. They claim that the results are similar to those found using a Heckman two stage procedure. The authors are concerned with how the returns might change in the future and they note that as more people acquire higher levels of education the return is likely to drop for those levels.

Chamberlain and van der Berg (2002) try to account for differences in the quality of education across race groups in determining the return to education. They proxy for quality using test scores from the PSLSD survey of 1993 and weight the years of schooling an individual has attained in the October Household Survey 1995 by using a predicted test score. Using a two stage selection procedure, they find that the return to education is around 5% before accounting for quality and that it increases to around 6% after accounting for quality.

Serumage-Zake and Naudé (2003) use a double hurdle model where they predict simultaneously whether a person will enter the labor market as well as whether they will find a job. They find returns around 12% using the 1995 October Household Survey.

Hertz (2003) shows that errors in the reporting of educational attainment can bias the estimated return to education upward. Using the PSLSD and KIDS panel data set in an OLS regression he shows that failing to correct for reporting error results in a return to education of 11 to 13%. Whereas, correcting for the error and using a within-family fixed effects approach reduces the return to between 5 and 6%. He shows that errors in the schooling variable are strongly correlated within families.

Keswell (2004), in a paper examining differences in the return to similar levels of schooling across race groups finds that the rate of return for Africans was around 11% at the end of Apartheid. This finding is from the PSLSD data set. He finds, using the Labour Force Surveys (LFS) of 2001 and 2002, that this return declined to 7%. These private returns to education are measured in OLS and Tobit regressions.

Keswell and Poswell (2004) use several data sets (PSLSD of 1993, October Household Survey (OHS)
of 1995 and 1997 and LFS of September 2000) to show that returns to higher levels of education in South Africa are convex. The estimation procedure they use is OLS allowing for non-linear returns to education in the form of polynomials in the second and third degree on the education variables. They find the return to primary school in 1993 is 2%, secondary school 28% and tertiary is between 68 and 72%. In 1995 the return to primary school decreases to zero, secondary school remains around 28% and tertiary education increases to between 71 and 86%. The return to secondary school decreases to 21% in 1997, and that to tertiary education to between 54 and 61%. Finally, in 2000, the return to primary education is negative, secondary drops to between 15 and 16% and tertiary remains constant. Leibbrandt, Levinsohn and McCrary (2005) use the October Household Survey and Income and Expenditure Survey of 1995 to compare South Africa’s income distribution to that found using the Labour Force Surveys and Income and Expenditure Survey in 2000. Using both descriptive methods and non-parametric techniques they find that the income distribution has shifted to the left over the five year period. In determining causes of the shift they find that the return to attributes has declined from 1995 to 2000. Specifically with respect to education, they find that the return to additional years of education decreased for African men and increased for white men. They claim that this result is expected due to continuing labor market rigidities. In 1995 for African men under 60 years of age, the return to education is between 11% and 14%. By 2000 the return for the youngest cohorts has declined by 4 percentage points. It remains constant for older cohorts.

Maitra and Vahid (2005) examine the effect of household characteristics on living standards between 1993 and 1998 in KwaZulu Natal. They use the KIDS panel data set. They account for non-random sample attrition since it appears that wealthier households were more likely to attrite. Using quantile regression techniques, they find that the return to education on log wage ranges from zero at the highest quantile and lowest level of education to 108% for the lowest quantile at the highest level of education. They find a negative correlation between education and ability, possibly a result of limited African access to occupations during apartheid. They find that by 1998 there is no longer any difference in the return across quantiles which they suggest is due to the openness of the labor market after the end of apartheid.

3. DERIVING MEANINGFUL AND FEASIBLE BOUNDS ON ATE

This section sets up the model and estimation framework and summarizes Manski and Pepper’s (2000) results. To do that, we need to introduce some notation now. A person has a realized treatment \( z \in T \) and a realized outcome \( y := y(z) \in Y \), both of which are observable. The outcome space has greatest lower bound \( K_0 := \inf Y \) and least upper bound \( K_1 := \sup Y \). The function \( y(\cdot) : T \to Y \) is called response function. The latent or conjectural outcomes \( y(t) \) (with \( t \in T \)) are not observed. Our goal is to come up with lower and upper bounds for \( E[y(t)|v] \), where \( v \), at this stage, is just some covariate (observed or unobserved). We are also interested in bounds on the unconditional mean \( E[y(t)] \).

Example: Our goal is to estimate the expected value of income (possibly conditional on \( v \)) for people
who go to school for 12 years. We are interested in \( E[y(12)|v] \). Why is this difficult? A naive idea is to just average the incomes of people who attended school for 12 years. Mathematically this approach estimates \( E[y|v, z = 12] := E[y(12)|v, z = 12] \neq E[y(12)|v] \). The problem, of course, is that we are dealing with a select sample; people have already sorted themselves into their preferred level of schooling.

3.1. No Assumptions Bound

Although the subsection title promises a bound without any assumptions, we will need to assume that the range of the dependent variable is bounded. The range of \( Y \) is \([K_0, K_1]\). The misnomer "no assumptions bound" is due to Manski (1989) and we do not intend to deviate from his convention. In any case, the assumption of bounded support will become redundant later; it only serves as an auxiliary assumption to derive the main results.

We start with the following decomposition of the conditional expectation:

\[
E[y(t)|v = u] = E[y(t)|v = u, z = t] \cdot \Pr(z = t|v = u) + E[y(t)|v = u, z \neq t] \cdot \Pr(z \neq t|v = u).
\]

Whenever the conjectural treatment \( t \) equals the actual treatment \( z \) we rewrite \( E[y(t)|v = u, z = t] = E[y|v = u, z = t] \) and thus

\[
E[y(t)|v = u] = E[y|v = u, z = t] \cdot \Pr(z = t|v = u) + E[y(t)|v = u, z \neq t] \cdot \Pr(z \neq t|v = u).
\]

This bit of rewriting helps see that the only term on the right hand side of the last equation that is not identifiable by the data is \( E[y(t)|v = u, z \neq t] \). What is the interpretation of \( E[y(t)|v = u, z \neq t] \)? It is the expected value of income after exactly \( t \) years of education for the subset of people who would choose to attend school for more or less than \( t \) years of schooling. It is the expected value of a counterfactual event. We do not observe this latent value in the data. The only thing we know about \( E[y(t)|v = u, z \neq t] \) is that it lies between \( K_0 \) and \( K_1 \). Therefore we use those extremes to bound the expectation:

\[
\begin{align*}
E[y(t)|v = u] &\geq E[y|v = u, z = t] \cdot \Pr(z = t|v = u) + K_0 \cdot \Pr(z \neq t|v = u) \\
E[y(t)|v = u] &\leq E[y|v = u, z = t] \cdot \Pr(z = t|v = u) + K_1 \cdot \Pr(z \neq t|v = u).
\end{align*}
\]

3.2. Monotone Instrumental Variable

The bounds in inequalities (3.1) are sharp in the absence of more information. But for practical purposes the no assumptions bounds have at least two drawbacks: They are likely not informative (if the range of \( Y \) is wide) and researchers usually do not know the values of \( K_0 \) and \( K_1 \). To ameliorate the first problem, Manski and Pepper (2000) introduce their monotone instrumental variable assumption.
Assumption 1 (Monotone Instrumental Variable)
Let \( V \) be an ordered set. The variable \( v \) is a monotone instrumental variable in the sense of mean-monotonicity if, for each \( t \in T \), and all \((u_1, u_2) \in (V \times V)\) such that \( u_2 \geq u_1 \)
\[
E[y(t)|v = u_2] \geq E[y(t)|v = u_1].
\]
How does this assumption help us? So far, the variable \( v \) did not play any active role in the derivation of the bounds. Now, if we observe different individuals with different values for \( v \) we can refine the bounds. By Assumption 1 we get, for all \( u_1 \leq u \leq u_2 \)
\[
\begin{align*}
E[y(t)|v = u] & \geq E[y(t)|v = u_1] \\
E[y(t)|v = u] & \leq E[y(t)|v = u_2].
\end{align*}
\]
For each of the expectations on the right hand side of the two inequalities (3.2), equation (3.1) provides bounds. The inequality \( E[y(t)|v = u] \geq E[y(t)|v = u_1] \) holds for all \( u_1 \leq u \). To tighten the bound on \( E[y(t)|v = u] \) we can pick the maximum value of the lower bound on \( E[y(t)|v = u_1] \), hence
\[
E[y(t)|v = u] \geq \sup_{u_1 \leq u} \{E[y|v = u_1, z = t] \cdot \Pr(z = t|v = u_1) + K_0 \cdot \Pr(z \neq t|v = u_1)\}.
\]
Likewise, to refine the upper bound on \( E[y(t)|v = u] \) we can pick the minimum value of the upper bound on \( E[y(t)|v = u_2] \), hence
\[
E[y(t)|v = u] \leq \inf_{u_2 \geq u} \{E[y|v = u_2, z = t] \cdot \Pr(z = t|v = u_2) + K_1 \cdot \Pr(z \neq t|v = u_2)\}.
\]
3.3. Monotone Treatment Selection
The assumption of monotone treatment selection is a modification of the MIV Assumption 1. When instrument and treatment coincide, \( v = z \), then the MIV Assumption collapses to

Assumption 2 (Monotone Treatment Selection)
Let \( T \) be an ordered set. For each \( t \in T \),
\[
u_2 \geq u_1 \Rightarrow E[y(t)|z = u_2] \geq E[y(t)|z = u_1].
\]
In order to derive the bounds on \( E[y(t)|v = u] = E[y(t)|z = u] \) we will use equations (3.3) and (3.4), the next proposition states the result.

Proposition 3 (MTS Bounds)
For a given latent treatment level \( t \in T \) and for different levels \( u \) of actual treatment \( z \), the bounds on \( E[y(t)|z = u] \) under Assumption 2 are
\[
\begin{align*}
u < t & \Rightarrow K_0 \leq E[y(t)|z = u] \leq E[y|z = t] \\
u = t & \Rightarrow E[y(t)|z = u] = E[y|z = t] \\
u > t & \Rightarrow E[y|z = t] \leq E[y(t)|z = u] \leq K_1.
\end{align*}
\]
Proof. See Appendix. ■
3.4. Monotone Treatment Response

The next step towards tightening the bounds on $E[y(t)|v = u]$ involves an assumption on the response function $y(\cdot): T \to Y$.

**Assumption 4 (Monotone Treatment Response)**

Let $T$ be an ordered set. For each $i \in I$, 
\[ t_2 \geq t_1 \Rightarrow y_i(t_2) \geq y_i(t_1). \]

Under the MTR Assumption Manski (1997, Corollary M1.2) derives bounds on $E[y(t)|v = u]$ that make more use of the information contained in the sample data than the MIV bounds (3.3) and (3.4).

**Proposition 5 (Monotone Treatment Response Bounds)**

The bounds on $E[y(t)|v = u]$ under Assumption 4 are
\begin{align*}
(3.5) \quad & E[y(t)|v = u] \geq E[y|v = u, z \leq t] \cdot \Pr(z \leq t|v = u) + K_0 \cdot \Pr(z > t|v = u) \\
(3.6) \quad & E[y(t)|v = u] \leq E[y|v = u, z \geq t] \cdot \Pr(z \geq t|v = u) + K_1 \cdot \Pr(z < t|v = u).
\end{align*}

**Proof.** See Appendix. ■

What is the advantage of the MTR bounds (3.5) and (3.6) over the MIV bound (3.3) and (3.4)? The MTR bounds use more information from the sample data. For example, the MTR lower bound uses the relatively uninformative range restriction $K_0$ on $Y$ only for treatment values $z > t$, for all values $z \leq t$ it exploits the more informative sample data. Compare this to the MIV lower bound which uses $K_0$ for all values of $z \neq t$ and exploits the sample data only for $z = t$. The MIV upper bound also makes productive use of the sample data only for $z = t$ while the MTR upper bound exploits all $z \geq t$. Altogether the MTR bounds thus use the entire range for $z$ while the MIV bounds only exploit the sample data for $z = t$.

3.5. Combining MIV with MTR

Our goal is to derive sharp bounds on $E[y(t)]$ without restricting the support $Y$ of $y(\cdot)$ to the interval $[K_0, K_1]$. Combining the MIV Assumption 1 with the MTR Assumptions 4 will yield bounds that still depend on $K_0$ and $K_1$, however the MIV-MTR bounds are the last crucial step that we need before we are ready to derive unrestricted bounds on $E[y(t)]$.

**Proposition 6 (MIV–MTR Bounds)**

The bounds on $E[y(t)|v = u]$ under Assumptions 1 and 4 are
\begin{align*}
(3.7) \quad & E[y(t)|v = u] \geq \sup_{u_1 \leq u} \{ E[y|v = u_1, z \leq t] \cdot \Pr(z \leq t|v = u_1) + K_0 \cdot \Pr(z > t|v = u_1) \}.
\end{align*}
\begin{align*}
(3.8) \quad & E[y(t)|v = u] \leq \inf_{u_2 \geq u} \{ E[y|v = u_2, z \geq t] \cdot \Pr(z \geq t|v = u_2) + K_1 \cdot \Pr(z < t|v = u_2) \}.
\end{align*}

**Proof.** See Appendix. ■

How do the MIV-MTR bounds (3.7) and (3.8) compare to the MIV bounds (3.3) and (3.4)? The MIV-MTR bounds make better use of the sample data than the MIV bounds. This is a direct consequence of the properties of the MTR bounds as discussed above in section 3.4.
3.6. Combining MTS with MTR

Now we are ready to derive bounds on \( E[y(t)] \) that do not rely on the artificial restriction that the support of \( y(\cdot) \) is bounded below by \( K_0 \) and above by \( K_1 \). Starting point is the observation that under MTS the instrument \( v \) and the treatment \( z \) coincide. In order to derive the bounds on \( E[y(t)|v = u] = E[y(t)|z = u] \) we will adapt equations (3.7) and (3.8) to the new environment in which \( v = z \). We need to be careful about how the actual treatment value \( u \) relates to the latent value \( t \). The discussion here is similar to section 3.3. The next proposition states the MTS-MTR bounds.

Proposition 7 (Sharp MTS-MTR Bounds on Conditional Expectation)

The bounds on \( E[y(t)|v = u] = E[y(t)|z = u] \) under Assumptions 2 and 4 are

\[
\begin{align*}
\mu(t) & \Rightarrow E[y|z = u] \leq E[y(t)|z = u] \leq E[y|z = t] \\
\mu(t) & \Rightarrow E[y|z = t] \leq E[y(t)|z = u] \leq E[y|z = t] \\
\Rightarrow E[y|z = t] = E[y(t)|z = u] \\
\mu(t) & \Rightarrow E[y|z = t] \leq E[y(t)|z = u] \leq E[y|z = u].
\end{align*}
\]

Proof. See Appendix. ■

To obtain bounds on the unconditional expectation, ‘integrate’ out over all values of \( z \):

Corollary 8 (Sharp MTS-MTR Bounds on Unconditional Expectation)

The bounds on \( E[y(t)] \) under Assumptions 2 and 4 are

\[
\begin{align*}
&\sum_{u < t} E[y|z = u] \cdot \Pr(z = u) + E[y|z = t] \cdot \Pr(z \geq t) \\
&\leq E[y(t)] \leq \sum_{u > t} E[y|z = u] \cdot \Pr(z = u) + E[y|z = t] \cdot \Pr(z \leq t).
\end{align*}
\]

Last, to put an upper bound on the average treatment effect \( \Delta(s, t) := E[y(t)] - E[y(s)] \), for \( s < t \), under both the MTR and MTS Assumptions, we merely add subtract the lower bound on \( E[y(s)] \) from the upper bound on \( E[y(t)] \):

\[
(3.9) \quad \Delta(s, t) \leq \sum_{u < s} \left( E[y|z = t] - E[y|z = u] \right) \cdot \Pr(z = u)
\]

\[
+ \left( E[y|z = t] - E[y|z = s] \right) \cdot \Pr(s \leq z \leq t)
\]

\[
+ \sum_{u > t} \left( E[y|z = u] - E[y|z = s] \right) \cdot \Pr(z = u).
\]

4. DATA

Ideally, we would like to follow the evolution of educational attainment in a panel data set. Unfortunately, the only panel data set available in South Africa focuses on only one of South Africa’s nine provinces. We therefore, conduct our analysis on both that panel as well as a country-wide cross-sectional data set to see both the evolution of the returns to education as well as to determine the return across all geographic regions.
TABLE I
—Data: Summary Statistics on Income, OHS/LFS and IES

<table>
<thead>
<tr>
<th></th>
<th>1995</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>23,432</td>
<td>22,470</td>
</tr>
<tr>
<td>s.d.</td>
<td>22,289</td>
<td>62,937</td>
</tr>
<tr>
<td>Skewness</td>
<td>4.01</td>
<td>51.53</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>42.10</td>
<td>3,448</td>
</tr>
<tr>
<td>Log Mean</td>
<td>9.68</td>
<td>9.49</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.94</td>
<td>1.08</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.46</td>
<td>0.75</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.98</td>
<td>5.49</td>
</tr>
<tr>
<td>Income percentile:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1,668</td>
<td>550</td>
</tr>
<tr>
<td>5</td>
<td>2,836</td>
<td>2,000</td>
</tr>
<tr>
<td>10</td>
<td>4,170</td>
<td>3,480</td>
</tr>
<tr>
<td>25</td>
<td>8,674</td>
<td>7,200</td>
</tr>
<tr>
<td>50</td>
<td>18,354</td>
<td>14,560</td>
</tr>
<tr>
<td>75</td>
<td>30,024</td>
<td>25,200</td>
</tr>
<tr>
<td>90</td>
<td>48,372</td>
<td>43,200</td>
</tr>
<tr>
<td>95</td>
<td>63,384</td>
<td>62,400</td>
</tr>
<tr>
<td>99</td>
<td>100,080</td>
<td>132,000</td>
</tr>
<tr>
<td>Observations</td>
<td>9,527</td>
<td>8,837</td>
</tr>
</tbody>
</table>


4.1. Labour Force Survey/October Household Survey and Income and Expenditure Survey

For the cross-sectional analysis, we use four data sets, two for 1995 and two for 2000 since there is no single data set that contains adequate information on both demographics and incomes. Since the two 1995 surveys interview the same households, we are able to match households across the surveys. This is true also for 2000, although there is no overlap of households between 1995 and 2000.

The 1995 data come from the October Household Survey (OHS95) and the Income and Expenditure Survey (IES95), both collected by Statistics South Africa. The OHS95 is one of a series of general household surveys run from 1994 intermittently until 1999 during October of each year. The questionnaires are not directly comparable over time, nor do the surveys interview the same households from year to year. The surveys contain information on household members’ demographic characteristics, education and health levels, employment status, access to infrastructure, and dwelling structures. The 1995 the survey interviewed 96,261 individuals from 29,700 households.

Since the OHS95 contains poor information on income, we supplement the data with income data from the IES95 which contains detailed information on household and personal income and expenditure. In 1995, the survey consists of 29,582 households and 128,917 individuals.\(^2\) The merged OHS95 and

---

\(^2\) As Leibbrandt, Levinsohn and McCrary (2005) point out, the reason the OHS95 income data is so poor might be because Statistics South Africa anticipated re-interviewing the same households for the IES95 later in the year.
IES95 data set consists of 114,568 individuals from 27,135 households.\(^3\) In addition, we inflate the data by 39%, the increase in the consumer price index over that period.

Turning to the 2000 data, the Labour Force Survey (LFS) replaced the OHS in 2000, and is implemented biannually once in March and once in September. We use the September 2000 survey as it largely consists of the same households as those in the IES 2000 (IES00). The questions in the LFS follow those in the OHS so that we are able to compare the OHS95 and LFS00 surveys. The LFS00 consists of 105,370 individuals from 26,571 households.

Since the IES00 asks the same questions as the IES95, they are also comparable. The IES00 consists of 104,153 individuals from 26,263 households. Upon merging the IES00 with the LFS00, we have 103,732 individuals from 26,150 households.\(^4\)

We restrict our analysis to African males between the ages of 18 and 65 who are employed and for whom we have education information which leaves us with 9,527 individuals in 1995 and 8,837 individuals in 2000. We ignore females to avoid potential complications with labor force and child rearing decisions. Our focus is on Africans as the population group most affected by Apartheid and post-Apartheid education policies.

Since our non-parametric analysis requires only income and education data, we provide summary statistics for these two variables in Table I and Table II. Table I shows that the mean real income of African males has declined between 1995 and 2000, consistent with the findings in Hoogeveen and Özler (2004) and Leibbrandt at el. (2005). Table II summarizes educational attainment in 1995 and 2000 by highest educational qualification obtained. A higher percentage of people completed high school in 2000 than in 1995, while a lower percentage obtained no education at all. Yet, overall, the aggregate educational attainment of employed African males hardly changed over the five year period.

\[4.2. \textit{KwaZulu-Natal Income Dynamics Study (KIDS)}\]

The KIDS panel data set provides more information on the evolution of the return to education than if we compare two cross-sectional samples. The first wave of the data set, conducted in 1993, was part of the World Bank’s Living Standard Measurement Study. The South African version, known as the Project for Statistics on Living Standards and Development is the first in the country to include the “independent” homelands, previously excluded from South African surveys for political reasons.\(^5\) It is therefore the first modern survey to provide some information on conditions in those former homeland areas. The survey covers all four races across the entire country.\(^6\) Furthermore it is the last survey to be taken prior to the historic democratic elections of 1994 and therefore provides a useful benchmark.

\(^3\)We drop people whose gender or race changes and those who age more than a year between surveys.

\(^4\)Again we delete any observations that do not match up such as people whose gender or race changed between the surveys or those who age more than a year.

\(^5\)The apartheid government created the homeland states as part of its separate development policy. The states were intended to provide citizenship to various African ethnicities and to be given independence. As such, individuals residing in these territories were not considered South African and were thus omitted from earlier surveys.

\(^6\)The four race groups are: African, White, Coloured, Indian.
TABLE II
—Data: Distribution of Education, OHS/LFS and IES

<table>
<thead>
<tr>
<th></th>
<th>1995</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>No schooling</td>
<td>14.74</td>
<td>11.03</td>
</tr>
<tr>
<td>Sub A/Sub B/Standard 1</td>
<td>5.88</td>
<td>7.11</td>
</tr>
<tr>
<td>Grade 4/Standard 2</td>
<td>6.46</td>
<td>5.24</td>
</tr>
<tr>
<td>Grade 5/Standard 3</td>
<td>6.08</td>
<td>5.57</td>
</tr>
<tr>
<td>Grade 6/Standard 4</td>
<td>7.43</td>
<td>6.79</td>
</tr>
<tr>
<td>Grade 7/Standard 5</td>
<td>8.60</td>
<td>8.98</td>
</tr>
<tr>
<td>Grade 8/Standard 6</td>
<td>10.35</td>
<td>9.27</td>
</tr>
<tr>
<td>Grade 9/Standard 7</td>
<td>6.46</td>
<td>6.87</td>
</tr>
<tr>
<td>Grade 10/Standard 8/NTCI</td>
<td>8.54</td>
<td>8.78</td>
</tr>
<tr>
<td>Grade 11/Standard 9/NTCH</td>
<td>4.59</td>
<td>6.53</td>
</tr>
<tr>
<td>Grade 12/Standard 10/NTCH</td>
<td>12.42</td>
<td>16.17</td>
</tr>
<tr>
<td>Diploma</td>
<td>6.42</td>
<td>5.38</td>
</tr>
<tr>
<td>Degree</td>
<td>2.05</td>
<td>2.29</td>
</tr>
</tbody>
</table>

Note.—In percent. Employed African males.

with which to compare subsequent survey findings and to evaluate the effect of post-Apartheid policies.

The second wave of the survey, the KwaZulu–Natal Income Dynamics Survey (KIDS), conducted in 1998, was directed by the University of Natal, the University of Wisconsin, and the International Food policy Research Institute. Due to financing constraints the second wave was limited to only one of South Africa’s nine provinces, KwaZulu–Natal, and surveys only the African and Indian households from the first wave. Coloured and white households are excluded because of the small sizes of these population groups in this province. Therefore the panel only exists in the one province for two race groups. Nevertheless, we believe the data offer valuable insights into changes in household characteristics and their impact on expenditure (income) particularly since the province now consists of both an old South African province (Natal) and a former homeland (KwaZulu).

In the interests of compatibility, the 1998 survey questionnaire largely replicates the 1993 questionnaire. An important aspect of the 1998 wave is that wherever possible enumerators tracked down households that had moved. The result is that 85% of the African and Indian households surveyed in KZN in 1993 were resurveyed in 1998. That is, 1178 of the original 1389 households were resurveyed. Maitra and Vahid (2006) show that sample attrition was not random and re–weight the second wave of the sample by the inverse probability that a household will attrite between 1993 and 1998. For the purposes of this study we use the already cleaned Maitra and Vahid data set available at http://qed.econ.queensu.ca/jae/2006-v21.7/maitra-vahid/. Upon merging the two waves, there are 1354 households in 1993 and 1132 households in 1998. After restricting the sample to African, male

---

7Details of the survey implementation can be read in May, Carter and Maluccio (2000).
8Accessed on February 6, 2009.
headed households there are 757 observations in 1993 and 557 in 1998. Table III summarizes household expenditure variables and household composition variables.

We are interested in the evolution of African returns to education as that is the group most disadvantaged by Apartheid education policies. Furthermore, we focus on male headed households to remove any gender based differences in the return to education. The household head education level increased slightly over the 5 year period with a higher percentage attending secondary school in 1998 than in 1993. However, the change is small, as Table IV shows.

Table V shows that household heads in most cases have completed their educational attainment by
the time they are the heads of a household. There is very little change over the time period apart from a
decrease in the percentage of household heads aged 75 and above with no education. This is mainly
due to an increase in the number of households with a younger head with education level 0.

5. ESTIMATION

5.1. Partially Identified Average Treatment Effects

The goal is to estimate the average treatment effect of education, \( \Delta(s, t) := E[y(t)] - E[y(s)] \), but
now we want to use less restrictive assumptions. We adapt Manski and Pepper’s (2000) partial identi-
fication method which has the advantage of more persuasive assumptions. This, of course, comes with
a disadvantage: The object of interest, \( \Delta(s, t) \), is not point identified. Instead, we will only be able to
bound it below a certain threshold. Nevertheless, a bounded estimate can be informative, which we
will see below.

For our estimation we only need two assumptions: The monotone treatment selection (MTS) Assump-
tion 2 and the monotone treatment response (MTR) Assumption 4.

What do these assumptions mean and how do they differ? As Manski and Pepper (2000) write, both
assumptions are distinct versions of the statement “wages increase with schooling.” The MTR Assump-
tion concerns the functional form of the income equation, it does not address the stochastic selection
process that makes people choose different levels of education. All it says is that more education will
weakly increase a person’s income, holding ability constant. The MTR Assumption deals with the di-
rect or pure effect that education has on earnings. The assumption does not deal with the indirect
effect that education could have through its correlation with ability (or any other covariates). This is
a statement regarding the (human capital production) functional form.

The MTS Assumption in contrast is concerned with the stochastic selection process that runs in the
background of the model. Schooling is an endogenous treatment, different people tend to select them-
selves into different education levels for a multitude of (often times unobserved) reasons. Consider the
following social experiment: You can force people to attend school for exactly \( t \) years. How would this
affect the subset of people who in the absence of your intervention would have chosen to attend school
for \( u_2 \) years? How would this affect the subset of people who in the absence of your intervention would
have chosen to attend school for \( u_1 < u_2 \) years? The MTS Assumption claims that you would expect
a higher income for the former group (i.e., the subset of people who chose \( u_2 \) years of education).

The MTR Assumption is consistent with the human capital accumulation model. The MTS Assump-
tion is weaker than a standard instrumental variable assumption and is consistent with the idea of
people with high ability selecting themselves into higher paying jobs. The validity of the MTR and
MTS Assumptions can be tested as proposed by Manski and Pepper (2000).

5.2. Validity of MTR and MTS Assumptions

Before presenting the estimator of the average treatment effect of education we clarify the mean-
ing of Assumptions 2 and 4 in the context of the sample data. As dependent variable we use per
capita household expenditures; as treatment variable we use years of education of the household head. The treatment variable is thus individual specific (household head) while the dependent variable is household specific. Maitra and Vahid (2006) follow the same approach, we thus analyze the average treatment effect of one additional year of education of the household head on the per capita expenditures of the household. How does this affect the interpretation of the MTR and MTS assumptions? Assumption 4 requires that an increase in the household head’s education weakly increases (ceteris paribus) household expenditures, which is a weak and reasonable assumption. The MTS assumption is more involved. Assumption 2 requires that the mean per capita expenditures of households with household heads that have high levels of education weakly exceed mean per capita expenditures of households with household heads that have lower levels of education. Is this assumption reasonable? Household heads with high levels of education tend to have higher ability and therefore earn a higher income (because ability is rewarded in the market as well), which in turn raises the per capita household expenditures of the high education households. Thus, the MTS assumption seems reasonable. On the other hand, households whose heads have low levels of education tend to be bigger, they include more working age adults and elderly (for whom households receive old age pensions). Households with more working age adults might have higher per capita household expenditures (although this is not necessary). There are thus two effects that go in different directions: an ability effect and a household size effect. A priori it is not clear which effect dominates. Ultimately it is a question of testing the MTR and MTS assumptions as proposed by Manski and Pepper (2000). As the results below show, we do not reject the MTR and MTS assumptions.

---

### TABLE V
---

<table>
<thead>
<tr>
<th>Education Level</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;75</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.127</td>
<td>0.221</td>
<td>0.251</td>
<td>0.485</td>
<td>0.567</td>
<td>0.690</td>
<td>0.316</td>
</tr>
<tr>
<td>I</td>
<td>0.692</td>
<td>0.355</td>
<td>0.392</td>
<td>0.531</td>
<td>0.402</td>
<td>0.333</td>
<td>0.276</td>
<td>0.415</td>
</tr>
<tr>
<td>II</td>
<td>0.308</td>
<td>0.518</td>
<td>0.363</td>
<td>0.190</td>
<td>0.114</td>
<td>0.078</td>
<td>0.034</td>
<td>0.254</td>
</tr>
<tr>
<td>III</td>
<td>0.000</td>
<td>0.000</td>
<td>0.025</td>
<td>0.028</td>
<td>0.000</td>
<td>0.022</td>
<td>0.000</td>
<td>0.016</td>
</tr>
</tbody>
</table>

### Year 1998
---

<table>
<thead>
<tr>
<th>Education Level</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;75</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.021</td>
<td>0.108</td>
<td>0.179</td>
<td>0.368</td>
<td>0.582</td>
<td>0.613</td>
<td>0.021</td>
<td>0.266</td>
</tr>
<tr>
<td>I</td>
<td>0.333</td>
<td>0.346</td>
<td>0.549</td>
<td>0.443</td>
<td>0.329</td>
<td>0.323</td>
<td>0.333</td>
<td>0.418</td>
</tr>
<tr>
<td>II</td>
<td>0.604</td>
<td>0.523</td>
<td>0.259</td>
<td>0.179</td>
<td>0.076</td>
<td>0.065</td>
<td>0.604</td>
<td>0.300</td>
</tr>
<tr>
<td>III</td>
<td>0.042</td>
<td>0.023</td>
<td>0.012</td>
<td>0.009</td>
<td>0.013</td>
<td>0.000</td>
<td>0.042</td>
<td>0.016</td>
</tr>
</tbody>
</table>

---

9 In 1998, for example, the correlation coefficient between years of education of the household head and number of working age adults in the household is -12.9%. The correlation coefficient between years of education of the household head and number of elderly in the household is -21.4%.
5.3. The Nonparametric Estimator

Equation (3.9) is the main result from section 3—the sharp upper bound on the average treatment effect of education. We want to express this upper bound explicitly in terms of further conditioning covariates $x$ to highlight the fact that our estimates below will depend on how we restrict the sample. Conditional on $x$, for every $s < t$, the upper bound becomes

$$
\Delta(s, t| x) := \sum_{u<s} \left( E[y|z = t, x] - E[y|z = u, x] \right) \Pr(z = u|x) \\
+ \left( E[y|z = t, x] - E[y|z = s, x] \right) \Pr(s \leq z \leq t|x) \\
+ \sum_{u>t} \left( E[y|z = u, x] - E[y|z = s, x] \right) \Pr(z = u|x).
$$

The conditioning variable $x$ defines our subpopulation of interest, it consists of people who are: male, black, and employed.

Last, we average out over time to obtain the average effect of one year of education: $\Delta(s, t|x)/(t - s)$.

5.4. The Modified Nonparametric Estimator

The KwaZulu-Natal Income Dynamics Study (KIDS) panel data set does not allow us to directly implement the Manski and Pepper estimator from equation (5.1). The problem is that we observe education only in four categorical variables: no education, primary school, some secondary school, and a high school degree. Table IV shows the definition of these categories.

Mechanically, we could simply replace $s$ and $t$ from equation (5.1) with the categories 0, I, II, and III. For example, we could compute an upper bound on $\Delta(I, II|x)$ via

$$
\Delta(I, II|x) \leq \left( E[y|z = II, x] - E[y|z = 0, x] \right) \Pr(z = 0|x) \\
+ \left( E[y|z = II, x] - E[y|z = I, x] \right) \Pr(I \leq z \leq II|x) \\
+ \left( E[y|z = III, x] - E[y|z = I, x] \right) \Pr(z = III|x).
$$

It is not clear at all, however, what the right hand side object identifies. A naive interpretation is that the right hand side is an upper bound on the average treatment effect of a category II education over a category I education. But both categories are so broad that no meaningful inferences can be made.

We are interested in bounding the average treatment effect of one additional year of education. It is not possible to discern this bound from the bound on $\Delta(I, II|x)$.

There exists one way, however, to compute a meaningful bound on the average treatment effect of one year of education. It involves the education categories 0 and II. The lowest category 0 has the advantage that, by definition, it only includes one realization of years of schooling, namely zero years of education. Category II ranges from 8 to 11 years of schooling. Comparing the (average) income of people from category II to the (average) income of people from category 0 is some sort of treatment effect of middle school versus no school at all. A conservative bound for the average treatment effect for people with 8 years of education (the minimum of category II) would be to say that all of the
income difference between the two categories can be attributed to the fact that everybody in category II must have had 8 years of education. This would surely overstate the treatment effect of receiving 8 years of education, but it could still yield a meaningful upper bound.\footnote{We would like to compare category 0 to category III as well, but this is not sensible here because of the small number of observations for category III.}

Ideally, of course, we would like to study the average treatment effect of one year of schooling, \( \Delta(0, 8) \),

\[
\Delta(0, 8) = \frac{1}{8} \sum_{u=9}^{12} \left( E[y|z = u, x] - E[y|z = 0, x] \right) \Pr(z = u|x).
\]

We are unable to compute a bound for \( \Delta(0, 8) \) because the education variable only exists in categories. We instead use the de-tour via

\[
\Delta(0, II|x) := \left( E[y|8 \leq z \leq 11, x] - E[y|z = 0, x] \right) \Pr(0 \leq z \leq 11|x) + \cdots
\]

\( (E[y|z = 12, x] - E[y|z = 0, x]) \Pr(z = 12|x) \),

and use it as a bound for \( \Delta(0, 8|x) \). The validity of using \( \Delta(0, II|x) \) as a bound for \( \Delta(0, 8|x) \) is shown by the next proposition.

**Proposition 9 (Informative Bound on ATE)**

\[
\Delta(0, 8|x) \leq \Delta(0, II|x).
\]

The proof is in the Appendix. To average it out over the 8 years we just compute \( \Delta(0, II|x)/8 \).

6. **RESULTS**

6.1. **IES and OHS/LFS Data**

Table VI presents, for the years 1995 and 2000, the empirical mean of log-income given years of schooling, \( \hat{E}[y|z = s] \), the distribution of years of education in the data, \( \hat{\Pr}(z = s) \), and the number of observations in each education category. (Recall that we restrict our sample to African males who are employed.) The columns for \( \hat{E}[y|z = s] \) and \( \hat{\Pr}(z = s) \) are the main ingredients for the estimation of the MTS-MTR bounds. What is striking in the South African case is the high number of uneducated \((s = 0)\) people: 1404 out of 8720 persons in 1995 and 976 out of 8179 persons in 2000 do not have any formal education. The low education level in South Africa is in stark contrast to developed countries. Manski and Pepper (2000), using the United States National Longitudinal Survey of Youth data set, have a minimum level of schooling of 8 years and a maximum level of schooling of 20 years. In the South African data, and with focus on African males who are employed, the maximum level of schooling is 12 years. We consider it advantageous, for bound estimation, that the minimum schooling level is zero. Our interest lies in comparing zero levels of schooling to higher levels of schooling. For example, we
compute $\Delta(0,8|x)/8$ to see what the benefits of eight years of education over no education are. Table VII shows that in the year 1995, $\Delta(0,8|x)/8$ the upper bound on the return was equal to 8.77%. This means that people with eight years of education have received an additional return of at most 8.77% per year on each of their eight years of education. We also provide the 95% quantile of the empirical distribution of our bound estimate. For $\Delta(0,8|x)/8$ the 95% quantile equals 9.41% which provides a more conservative evaluation of the upper bound.

For the year 1995, the bound estimate is more or less stable for $t = 8, 9, 10, 11$. When $t = 12$, which represents the final year of high school, the bound increases to 10.23%. In all cases, the MTS-MTR bound lies below the OLS estimator of 10.36%. To obtain this estimate we run a typical Mincerian wage regression of log-income on a constant, education, age (linear and quadratic), and province (there are nine provinces in the data). The OLS estimator is, of course, biased. Card (2001) comprehensively discusses different sources of bias in the OLS estimator of the return to schooling. Summarizing research findings in the area, Card argues that, if anything, the OLS estimator is downwards biased (for example, comparing to instrumental variables estimation of discrete dynamic choice programming model estimation). We therefore regard the OLS estimator as a lower bound for parametric estimation. In this sense then the bound estimates $\Delta(0,t|x)/t$ reported in Table VII for the year 1995 are informative because they fall below estimates that result from parametric estimation. Even if we base our conclusions on only the conservative 95% quantile, the bound estimates for $t = 8, 9, 10, 11$ are all lower than parametric estimates. For the year 2000 the bound results are even stronger. For all

---

### Table VI

| s  | $E[y|z=s]$ | $Pr(z=s)$ | Sample Size | $E[y|z=s]$ | $Pr(z=s)$ | Sample Size |
|----|------------|------------|-------------|------------|------------|-------------|
| 0  | 8.66       | 0.161      | 1404        | 8.89       | 0.119      | 976         |
| 1  | 8.63       | 0.010      | 80          |            |            |             |
| 2  | 8.59       | 0.064      | 560         | 8.76       | 0.020      | 163         |
| 3  | 8.72       | 0.071      | 615         | 9.11       | 0.057      | 463         |
| 4  | 8.71       | 0.066      | 579         | 9.05       | 0.060      | 493         |
| 5  | 8.86       | 0.081      | 708         | 9.19       | 0.074      | 601         |
| 6  | 9.02       | 0.094      | 819         | 9.25       | 0.097      | 795         |
| 7  | 9.22       | 0.113      | 986         | 9.37       | 0.101      | 822         |
| 8  | 9.27       | 0.071      | 615         | 9.33       | 0.075      | 610         |
| 9  | 9.51       | 0.093      | 814         | 9.49       | 0.095      | 778         |
| 10 | 9.58       | 0.050      | 437         | 9.42       | 0.071      | 577         |
| 11 | 9.88       | 0.136      | 1183        | 9.72       | 0.175      | 1432        |
| Total | 1         | 8720       | 1           | 1          | 8179       |             |

Note.—In 1995 years of schooling $s = 1, 2, 3$ are comprised in one category. Tertiary education has been removed due the lack of clarity about how many years of education a tertiary qualification entails.


6.2. KIDS Data

We have computed upper bounds for the average treatment effect of one year of schooling based on \( \frac{\Delta(0, s|x)}{s} \) via \( \Delta(0, II|x) \) as explained above. The results are in Table VIII.

The average treatment effect of one additional year of education in 1993 is below 7.1%. For the year 1998 the upper bound is 8.7%. For the bound estimates we obtain bootstrap standard errors of 0.0077 and 0.0087 for 1993 and 1998. These standard errors result from the empirical distribution of the data that we obtained by simulating 5,000 bootstrap repetitions. The numbers for the 95% quantiles are also derived from the empirical distribution, 95% of the bound estimates will fall below the reported quantiles of 8.30% and 10.14% for 1993 and 1998.

The fact that the bound estimate in 1993 lies below the 1998 estimate, of course, does not imply that the average treatment effect has increased from 1993 to 1998. By the nature of our bound estimator, the average treatment effect is not point identified. The bounds on the average treatment effects are lower than most of the parametric estimates from the literature. For example, Mwabu and Schultz (2000) report annual returns to education of 15.8% for African men with secondary education in the year 1993. Our results suggest that their estimate is severely upward biased—by more than 100%.
7. CONCLUSION

We non-parametrically estimate upper bounds on the average treatment effect of one additional year of schooling in South Africa for two different types of data during the 1990s. Using cross-sectional data from combining the OHS95 and IES95 in 1995 and LFS00 and IES00 in 2000 we find an upper bound on the return to education of 8.77% in 1995 and 6.95% in 2000. In contrast, using a panel data set that covers only one of South Africa’s nine provinces, we find an upper bound on the return to education of 7.11% in 1993 and 8.72% in 1998. Compared to existing parametric estimates in the literature, our bound is informative. We do not find much variation between the cross-sectional sample and the panel sample.

We anticipate that any parameter bound can be made tighter by imposing additional assumptions. We already achieve meaningful bounds using only the mild MTR and MTS conditions of Manski and Pepper (2000). However, we plan to extend our work along the lines of Blundell et al. (2007), who derive a nonparametric bound that accounts for sample selection into occupations. They do this by adding a stochastic dominance type condition.
APPENDIX A: PARAMETRIC IDENTIFICATION OF ATE

A.1. The Human Capital Earnings Function: Schooling and Ability

Our goal is to estimate the causal effect \( \text{ceteris paribus} \) of education on earnings. To do that, assume that there exists a human capital production function for individual \( i \)

\[
y_i(s, a) : T \times A \rightarrow Y
\]

that maps years of schooling \( s \in T \) and ability \( a \in A \) into individual earnings outcomes \( y_i(s, a) \in Y \). This function is a smaller version of Mincer’s (1974) human capital earnings function. Manski (1997) calls equation (A.1) a response function. The function simply illustrates that schooling \( s \) and ability \( a \) have a direct or pure effect on earnings.

For practical purposes the presence of ability in the response function creates at least three problems:

(i) Ability is not well defined.

(ii) Ability is not measured with sufficient accuracy.

(iii) Ability is not part of most data sets, and hence unobserved.

Standard cross-sectional data sets typically collect years of education, \( s \), along with income \( y \). Estimation has to rely on these variables only. Going back to equation (A.1) we define

\[
y_i(\cdot, \bar{a}) : T \rightarrow Y,
\]

which is a univariate function (holding ability constant) mapping schooling into earnings outcomes. Our goal is to measure the pure effect of schooling on earnings \( \text{ceteris paribus} \):

\[
\Delta(s, t) := E[y(t, \bar{a})] - E[y(s, \bar{a})]
\]

where \( s \in T \) and \( t \in T \) are years of education and \( s < t \). The object \( \Delta(s, t) \) is the average treatment effect of schooling on earnings, the estimation of which is our research objective.

A.2. Digression: Naive Nonparametric Estimation

In the sample data we observe, for every person \( i \), her level of schooling, \( z_i \in T \), and her income \( y_i := y_i(z_i) \in Y \). Note that we distinguish between the latent levels of schooling, denoted \( s \) and \( t \), and the observed level of schooling for person \( i \), denoted \( z_i \). The latent outcomes \( y_i(s) \) or \( y_i(t) \) are not observed whenever \( s \neq z_i \) or \( t \neq z_i \). To estimate the effect of schooling on earnings we use a random sample of people for which we observe data pairs on schooling and income \( \{(z_i, y_i) \in T \times Y\} \). A naive nonparametric solution for estimating \( \Delta(s, t) \) is to simply average those income observations \( y_i \) for which \( z_i = s \), and compare them to the average of the income observations for which \( z_i = t \). The problem with this approach is that ability and schooling are correlated. The observed data are realizations of peoples’ optimization decisions in which ability can be seen as a state variable and schooling as a choice variable.\(^{11}\) High ability people are more likely to choose higher levels of schooling (and vice versa). Schooling is hence endogenous, the resulting estimator for \( \Delta(s, t) \) is biased upward.

A.3. Problems of Parametric Estimation

There are at least two problems with parametric estimation: functional form and selection. Writing log-earnings as a linear function of schooling looks simplistic but it is convention in labor economics. Card (2001) argues for an additional quadratic schooling term so that the marginal effect of schooling is declining in schooling (assuming that the coefficient of the quadratic schooling term is negative). But even this assumption seems arbitrary. Because the relationship between earnings and schooling is not governed by a deterministic law, there will always remain different opinions about functional form.

\(^{11}\)Keane and Wolpin (1997) develop a dynamic choice programming model along those lines.
Regarding selection, the OLS model simply disregards the problem of correlation between schooling and ability. Consider two persons with different levels of ability. If in a social experiment we could force both of them to obtain the same amount of schooling $t$ then disregarding any selection we would expect both of them to have the same income:

\[(A.3) \quad E[y_i(t)|a_i = a_1] = E[y_i(t)|a_i = a_2] \quad \text{for } a_1 \neq a_2.\]

It is more realistic, however, to think that the person with higher ability would receive a higher income, in which case the equality in equation (A.3) turns into an inequality. The availability of an instrumental variable changes this interpretation a bit. Equation (A.3) is replaced by

\[E[y_i(t)|q_i = q_1] = E[y_i(t)|q_i = q_2],\]

for $q_1 \neq q_2$. This claim holds by the definition of an instrumental variable.

### A.4. Ordinary Least Squares Estimation

The first assumption of any parametric analysis always is linearity. Countless papers in labor economics run versions of the following regression:

\[(A.4) \quad y_i = z_i \beta + a_i + \varepsilon_i,\]

where $y_i$ is the logarithm of earnings, $a_i$ is unobserved ability, $\varepsilon_i$ is a random error, and $\beta$ is some coefficient. Equation (A.4) is a simplification of Mincer’s (1974) human capital earnings function, with two essential features: the linear link between schooling and log–earnings and the effect of unobserved ability on earnings. Using a random sample $\{z_i, y_i\}$, the classical linear regression model simply estimates $\beta$ as the slope coefficient using the assumption

\[(A.5) \quad E[a_i + \varepsilon_i|z_i] = 0,\]

Under equations (A.4) and (A.5) the average treatment effect $\Delta(s, s + 1)$ equals

\[\Delta(s, s + 1) = E[y_i(s + 1)] - E[y_i(s)] = (z_i + 1)\beta - z_i\beta = \beta.\]

To estimate the average treatment effect we therefore only need to run an OLS regression and obtain the slope coefficient.

### A.5. Instrumental Variables Estimation

Parametric estimation seems like a convenient way to estimate $\Delta(s, s + 1)$. And given the assumptions so far, it is also the best linear unbiased estimator. An obvious drawback is assumption (A.5). A better set of assumptions would be

\[E[a_i|z_i] \neq 0, \quad E[\varepsilon_i|z_i] = 0.\]

Ability is unobserved in the data. Running an ordinary least squares regression of log–earnings on education yields an inconsistent estimate for $\beta$. The average treatment effect is not identified. The way around this problem is instrumental variables estimation. In order to identify $\beta$ we need an instrumental variable, $v \in V$, which satisfies:

(i) Constant treatment response: $y(s, v_1) = y(s, v_2) = y(s)$ for all $v_1 \neq v_2$

(ii) Correlation: $E[s_i|v_i] = \pi v_i$ with $\pi \neq 0$

(iii) Exogeneity: $E[a_i|v_i] = 0$.

\[12\]The term ‘constant treatment response’ was first defined by Manski and Pepper (2008).
The average treatment effect equals
\[ \Delta(s, s + 1) = E[y_i(s + 1)] - E[y_i(s)] \]
\[ = E[(s_i + 1)\beta + a_i + \varepsilon_i|s_i] - E[s_i\beta + a_i + \varepsilon_i|s_i] \]
\[ = E[(s_i + 1)\beta + a_i + \varepsilon_i|\varepsilon_i] - E[s_i\beta + a_i + \varepsilon_i|\varepsilon_i] \]
\[ = \beta. \]

**APPENDIX B: MATHEMATICAL PROOFS**

**Proof of Proposition 3.** We need to be careful about how the actual treatment value \( u \) relates to the latent treatment value \( t \). There are three cases.

(i) If \( u < t \) then the term \( E[y|v = u_1, z = t] \) in equation (3.3) equals \( E[y|z = u_1, z = t] \) which is undefined for \( u_1 \leq u \). The probability \( \Pr(z = t|v = u_1) = \Pr(z = t|z = u_1) = 0 \) for \( u_1 \leq u \). The only thing remaining for the lower bound is \( K_0 \), noting that \( \Pr(z \neq t|v = u_1) = \Pr(z \neq t|z = u_1) = 1 \) for \( u_1 \leq u \). Hence, if \( u < t \) then \( K_0 \leq E[y(t)|z = u] \). Next, to derive the upper bound when \( u < t \) consider equation (3.4). How do we find the infimum? We have the degree of freedom to set \( u_2 = t \) which gives an upper bound of \( E[y(t)|z = t] = E[y|z = t] \). For all other values of \( u_2 \geq t \) the upper bound will just be \( K_1 \) which exceeds \( E[y|z = t] \). Thus, if \( u < t \) then \( E[y(t)|z = t] =: E[y|z = t] \). The probability in the derivation of the lower bound from equation (3.3) we have the degree of freedom to set \( y_i = z \) which is less than \( t \) then the supremum of the lower bound in equation (3.3) is \( y_i \). The probability \( \Pr(z = t|v = u_2) = \Pr(z = t|z = u_2) = 0 \) for \( u_2 \geq u \). The only thing remaining for the upper bound is \( K_1 \), noting that \( \Pr(z \neq t|v = u_2) = \Pr(z \neq t|z = u_2) = 1 \) for \( u_2 \geq u \). Hence, if \( u > t \) then \( K_1 \geq E[y(t)|z = u] \).

(ii) If \( u = t \) then the supremum of the lower bound in equation (3.3) is \( E[y(t)|z = t] \) which equals \( E[y|z = t] \). The infimum of the upper bound is also equal to \( E[y|z = t] \).

(iii) If \( u > t \) in the derivation of the lower bound from equation (3.3) we have the degree of freedom to set \( u_1 = t \) which gives a lower bound of \( E[y(t)|z = t] =: E[y|z = t] \). For all other values of \( u_1 \leq u \) the upper bound will just be \( K_0 \) which is less than \( E[y|z = t] \). Thus, if \( u > t \) then \( E[y(t)|z = t] =: E[y|z = t] \leq E[y(t)|z = u] \). Next, to derive the upper bound when \( u > t \) consider equation (3.4). How do we find the infimum? If \( u > t \) then the term \( E[y|v = u_2, z = t] \) in equation (3.4) equals \( E[y|z = u_2, z = t] \) which is undefined for \( u_2 \geq u \). The probability \( \Pr(z = t|v = u_2) = \Pr(z = t|z = u_2) = 0 \) for \( u_2 \geq u \). The only thing remaining for the upper bound is \( K_1 \), noting that \( \Pr(z \neq t|v = u_2) = \Pr(z \neq t|z = u_2) = 1 \) for \( u_2 \geq u \). Hence, if \( u > t \) then \( K_1 \geq E[y(t)|z = u] \).

In summary, we have
\[ u < t \Rightarrow K_0 \leq E[y(t)|z = u] \leq E[y(t)|z = t] =: E[y|z = t] \]
\[ u = t \Rightarrow E[y(t)|z = u] = E[y|z = t] \]
\[ u > t \Rightarrow E[y(t)|z = t] =: E[y|z = t] \leq E[y(t)|z = u] \leq K_1. \]

**Proof of Proposition 5.** By the MTR Assumption 4 we have
\[ t < z \Rightarrow K_0 \leq y(t) \leq y(z) \]
\[ t = z \Rightarrow y(t) = y(z) \]
\[ t > z \Rightarrow y(z) \leq y(t) \leq K_1. \]

This carries over immediately to the conditional expectations:

(B.1) \[ t < z \Rightarrow K_0 \leq E[y(t)|v = u] \leq E[y(z)|v = u, z > t] =: E[y|v = u, z > t] \]

(B.2) \[ t = z \Rightarrow E[y(t)|v = u] = E[y(z)|v = u, z = t] =: E[y|v = u, z = t] \]

(B.3) \[ t > z \Rightarrow E[y(z)|v = u, z < t] =: E[y|v = u, z < t] \leq E[y(t)|v = u] \leq K_1. \]

Now we combine the two lower bounds from equations (B.1) and (B.3) by weighting them with their probabilities, \( P(t < z|v = u) \) and \( P(t > z|v = u) \), and we also include equation (B.2) to obtain the MTR lower bound
\[ E[y(t)|v = u] \geq K_0 \cdot \Pr(z > t|v = u) + E[y(z)|v = u, z \leq t] \cdot \Pr(z \leq t|v = u). \]
Likewise, we combine the two upper bounds from equations (B.1) and (B.3) by weighing them with their probabilities, \( P(t < z|v = u) \) and \( P(t > z|v = u) \), and we also involve equation (B.2) to obtain the MTR upper bound

\[
E[y(t)|v = u] \leq E[y(z)|v = u, z \geq t] \cdot \Pr(z \geq t|v = u) + K_1 \cdot \Pr(z < t|v = u).
\]

\[
\text{Proof of Proposition 6.} \quad \text{The MIV-MTR bounds follow immediately by adapting the derivation of the MIV bounds in section 3.2. Starting point is Assumption 1. We get, for all } u_1 \leq u \leq u_2
\]

(B.4) \[
E[y(t)|v = u] \geq E[y(t)|v = u_1]
\]

\[
E[y(t)|v = u] \leq E[y(t)|v = u_2].
\]

For each of the expectations on the right hand side of the two inequalities (B.4), equations (3.5) and (3.6) provide bounds. The inequality \( E[y(t)|v = u] \leq E[y(t)|v = u_2] \) holds for all \( u_1 \leq u \). To tighten the bound on \( E[y(t)|v = u] \) we can pick the maximum value of the lower bound (3.5), hence

(B.5) \[
E[y(t)|v = u] \geq \sup_{u_1 \leq u} \{ E[y|v = u_1, z \leq t] \cdot \Pr(z \leq t|v = u_1) + K_0 \cdot \Pr(z > t|v = u_1) \}.
\]

Likewise, to refine the upper bound on \( E[y(t)|v = u] \) we can pick the minimum value of the upper bound (3.6), hence

(B.6) \[
E[y(t)|v = u] \leq \inf_{u_2 \geq u} \{ E[y|v = u_2, z \geq t] \cdot \Pr(z \geq t|v = u_2) + K_1 \cdot \Pr(z < t|v = u_2) \}.
\]

\[
\text{Proof of Proposition 7.} \quad \text{There are three cases.}
\]

(i) If \( u < t \) the lower bound will be the supremum (across \( u_1 \)) of \( E[y|z = u_1, z \leq t] \cdot \Pr(z \leq t|z = u_1) + K_0 \cdot \Pr(z > t|z = u_1) \). The term \( \Pr(z > t|z = u_1) = 0 \) while \( \Pr(z \leq t|z = u_1) = 1 \) so the supremum has to be based on \( E[y|z = u_1, z \leq t] \). Therefore, the lower bound is \( \sup_{u_1 \leq u} \{ E[y|z = u_1, z \leq t] \} \) which is equal to \( \sup_{u_1 \leq u} \{ E[y|z = u_1] \} \) when \( u < t \). Contrast the lower bound here to the one in section 3.3 for \( u < t \). Imposing the MTS assumption only, the lower bound was \( K_0 \). Already it emerges that the lower bound under MTS-MTR is more informative than under MTS only. As for the upper bound, the infimum of \( E[y|z = u_2, z \geq t] \cdot \Pr(z \geq t|z = u_2) + K_1 \cdot \Pr(z < t|z = u_2) \) will be based on \( E[y|z = u_2, z \geq t] \). Therefore, the upper bound for the case \( u < t \) equals \( \inf_{u_2 \geq t} \{ E[y|z = u_2, z \geq t] \} \) which is equal to \( \inf_{u_2 \geq t} \{ E[y|z = u_2] \} \). (Note that we use \( \inf_{u_2 \geq t} \) and not \( \inf_{u_2 \geq u} \). This holds because the intersection of \( \{ z = u_2 \} \) and \( \{ z \geq t \} \) when \( u < t \) and for all \( u_2 \geq u \) equals \( \{ z \geq t \} \).

(ii) If \( u = t \) then the supremum of the lower bound in equation (3.7) is \( \sup_{u_1 \leq u} \{ E[y|z = u_1] \} \). The infimum of the upper bound is also equal to \( \inf_{u_2 \geq t} \{ E[y|z = u_2] \} \) which is the same as for the case \( u < t \).

(iii) If \( u > t \) the lower bound will be the supremum (across \( u_1 \)) of \( E[y|z = u_1, z \leq t] \cdot \Pr(z \leq t|z = u_1) + K_0 \cdot \Pr(z > t|z = u_1) \). The term \( \Pr(z > t|z = u_1) = 0 \) while \( \Pr(z \leq t|z = u_1) = 1 \) so the supremum has to be based on \( E[y|z = u_1, z \leq t] \). Therefore, the lower bound is \( \sup_{u_1 \leq t} \{ E[y|z = u_1, z \leq t] \} \) which is equal to \( \sup_{u_1 \leq u} \{ E[y|z = u_1] \} \) when \( u < t \). (Note that we use \( \sup_{u_1 \leq t} \) and not \( \sup_{u_1 \leq u} \). This holds because the intersection of \( \{ z = u_1 \} \) and \( \{ z \leq t \} \) when \( u > t \) and for all \( u_1 \leq u \) equals \( \{ z \leq t \} \).) As for the upper bound, the infimum of \( E[y|z = u_2, z \geq t] \cdot \Pr(z \geq t|z = u_2) + K_1 \cdot \Pr(z < t|z = u_2) \) will be based on \( E[y|z = u_2, z \geq t] \). Therefore, the upper bound for the case \( u > t \) equals \( \inf_{u_2 \geq u} \{ E[y|z = u_2, z \geq t] \} \) which is equal to \( \inf_{u_2 \geq u} \{ E[y|z = u_2] \} \). Contrast the upper bound here to the one in section 3.3 for \( u > t \). Imposing the MTS assumption only, the upper bound was \( K_1 \). Again it is true that the bound under MTS-MTR is more informative than under MTS only.

In summary, we have

(B.7) \[
\begin{align*}
 u < t & \Rightarrow \sup_{u_1 \leq u} \{ E[y|z = u_1] \} \leq E[y(t)|z = u] \leq \inf_{u_2 \geq t} \{ E[y|z = u_2] \} \\
 u = t & \Rightarrow \sup_{u_1 \leq u} \{ E[y|z = u_1] \} \leq E[y(t)|z = u] \leq \inf_{u_2 \geq t} \{ E[y|z = u_2] \} \\
 u > t & \Rightarrow \sup_{u_1 \leq t} \{ E[y|z = u_1] \} \leq E[y(t)|z = u] \leq \inf_{u_2 \geq u} \{ E[y|z = u_2] \}.
\end{align*}
\]
Next, note that
\begin{equation}
B.8 \quad u_1 \leq u_2 \Rightarrow E[y|z = u_1] := E[y(u_1)|z = u_1] \quad \text{(definition)}
\leq E[y(u_2)|z = u_1] \quad \text{(by MTR)}
\leq E[y(u_2)|z = u_2] \quad \text{(by MTS)}
\leq E[y|z = u_2] \quad \text{(definition)}.
\end{equation}

The observation that \(E[y|z = u_1] \leq E[y|z = u_2]\), which may appear obvious, does not follow from the MTR Assumption or from the MTS Assumption alone, but only from combining the two. The result that \(E[y|z = u_1] \leq E[y|z = u_2]\) implies that

\[
\begin{align*}
\sup_{u_1 \leq u} \{E[y|z = u_1]\} &= E[y|z = u] \\
\inf_{u_2 \geq t} \{E[y|z = u_2]\} &= E[y|z = t] \\
\sup_{u_1 \leq t} \{E[y|z = u_1]\} &= E[y|z = t] \\
\inf_{u_2 \geq u} \{E[y|z = u_2]\} &= E[y|z = u],
\end{align*}
\]

and therefore by combining the results from equations (B.7) and equations (B.8) we obtain informative bounds that do not rely on bounded support \(Y\) of \(y(\cdot)\):

\[
\begin{align*}
u < t &\Rightarrow E[y|z = u] \leq E[y(t)|z = u] \leq E[y|z = t] \\
u = t &\Rightarrow E[y|z = t] \leq E[y(t)|z = u] \leq E[y|z = t] \\
&\Rightarrow E[y|z = t] = E[y(t)|z = u] \\
u > t &\Rightarrow E[y|z = t] \leq E[y(t)|z = u] \leq E[y|z = u].
\end{align*}
\]

\textbf{Proof of Proposition 9.} We suppress the conditioning on \(x\) to cut down notation. Define

\[
\begin{align*}
\lambda_1 := &\ (E[y|z = 8] - E[y|z = 0]) \Pr(0 \leq z \leq 8) \\
\lambda_2 := &\sum_{u=9}^{11} (E[y|z = u] - E[y|z = 0]) \Pr(z = u) \\
\lambda := &\lambda_1 + \lambda_2,
\end{align*}
\]

and obtain the lower bound in equation (5.2) as \(\lambda = (E[y|z = 12] - E[y|z = 0]) \Pr(z = 12)\). A sufficient condition to establish the proposition is

\begin{equation}
B.9 \quad \lambda \leq (E[y|8 \leq z \leq 11] - E[y|z = 0]) \Pr(0 \leq z \leq 11).
\end{equation}

For the right hand side in equation (B.9) we get

\[
\begin{align*}
(E[y|8 \leq z \leq 11] - E[y|z = 0]) \Pr(0 \leq z \leq 11)
&= \Pr(0 \leq z \leq 11) \sum_{u=8}^{11} (E[y|z = u] - E[y|z = 0]) \Pr(z = u) \\
&= \Pr(0 \leq z \leq 11) \Pr(8 \leq z \leq 11) \left[ \sum_{u=8}^{11} (E[y|z = u] - E[y|z = 0]) \Pr(z = u) \right] \\
&= \Pr(0 \leq z \leq 11) \Pr(z = 8) \left( E[y|z = 8] - E[y|z = 0] \right) + \cdots \\
&\quad \sum_{u=9}^{11} (E[y|z = u] - E[y|z = 0]) \Pr(z = u) + \cdots \\
&\quad \Pr(0 \leq z \leq 11) \Pr(8 \leq z \leq 11) \sum_{u=9}^{11} (E[y|z = u] - E[y|z = 0]) \Pr(z = u) + \cdots \\
&\quad \sum_{u=9}^{11} (E[y|z = u] - E[y|z = 0]) \Pr(z = u).
\end{align*}
\]
To prove the proposition we now only need to show that
\[
\lambda_1 \leq \frac{\Pr(0 \leq z \leq 11) \Pr(z = 8)}{\Pr(8 \leq z \leq 11)} \left( E[y|z = 8] - E[y|z = 0] \right) + \cdots
\]
\[
\left[ \frac{\Pr(0 \leq z \leq 11)}{\Pr(8 \leq z \leq 11)} - 1 \right] \sum_{u=9}^{11} \left( E[y|z = u] - E[y|z = 0] \right) \Pr(z = u)
\]
\[
= \Gamma.
\]
Using monotonicity, we can bound the right hand side:
\[
\Gamma \geq \frac{\Pr(0 \leq z \leq 11) \Pr(z = 8)}{\Pr(8 \leq z \leq 11)} \left( E[y|z = 8] - E[y|z = 0] \right) + \cdots
\]
\[
\left[ \frac{\Pr(0 \leq z \leq 11)}{\Pr(8 \leq z \leq 11)} - 1 \right] \left( E[y|z = 8] - E[y|z = 0] \right) \Pr(9 \leq z \leq 11)
\]
\[
= \frac{E[y|z = 8] - E[y|z = 0]}{\Pr(8 \leq z \leq 11)} \left[ \Pr(9 \leq z \leq 11) \Pr(0 \leq z \leq 7) + \Pr(z = 8) \Pr(0 \leq z \leq 11) \right]
\]
\[
= \frac{E[y|z = 8] - E[y|z = 0]}{\Pr(8 \leq z \leq 11)} \left[ \Pr(9 \leq z \leq 11) \Pr(0 \leq z \leq 7) + \cdots \right]
\]
\[
\Pr(z = 8) \Pr(0 \leq z \leq 8)
\]
\[
= \frac{E[y|z = 8] - E[y|z = 0]}{\Pr(8 \leq z \leq 11)} \left[ \Pr(9 \leq z \leq 11) \Pr(0 \leq z \leq 8) + \Pr(z = 8) \Pr(0 \leq z \leq 8) \right]
\]
\[
= \left( E[y|z = 8] - E[y|z = 0] \right) \frac{\Pr(0 \leq z \leq 8)}{\Pr(8 \leq z \leq 11)} \left[ \Pr(9 \leq z \leq 11) + \Pr(z = 8) \right]
\]
\[
= \left( E[y|z = 8] - E[y|z = 0] \right) \Pr(0 \leq z \leq 8)
\]
\[
= \lambda_1.
\]

REFERENCES


