Estimating the Return to Endogenous Schooling Decisions via Conditional Second Moments

Roger Klein
Francis Vella

ABSTRACT
This paper employs conditional second moments to identify the impact of education in wage regressions where education is treated as endogenous. This approach avoids the use of instrumental variables in a setting where instruments are frequently not available. We employ this methodology to estimate the returns to schooling for a sample of Australian workers. We find that accounting for the endogeneity of education in this manner increases the estimated return to education from 6 percent to 10 percent.

I. Introduction

The impact of education on earnings has important implications for education policy and individual investment decisions in human capital. Accordingly, estimating this impact has become an important objective in empirical labor economics. However, the level of education is generally chosen by the individual and unobserved factors that influence this choice, such as ability and motivation, are also likely to have a direct effect on earnings. Due to this potential simultaneity, accounting for the endogeneity of schooling is an integral part of estimating how earnings respond to educational investments (for a discussion, see Card 2001).

Recent innovations in this literature have generally focused on instrumental variables (IV) estimation. One popular approach is the creative construction of variables...
to employ as instruments for schooling. A well-known example is the use of a "policy shock" such as a change in the compulsory schooling laws (see, for example, Angrist and Krueger 1991). A second popular approach is to exploit repeated measurements on the same units of observation to eliminate unobserved heterogeneity via appropriate data transformations. This includes the use of panel data where the basis of the unobserved heterogeneity is the individual (for example, Griliches 1978) and family or twins data where the "family effect" is seen as the source of the unobserved heterogeneity (see, for example, Ashenfelter and Krueger 1994). Note, that it is well known that estimators which employ such transformations also have an IV interpretation.

These IV type procedures are based on orthogonality conditions involving the sample first moments of wages and education or the sample moments of differences in wages and education. For example, when one uses multiple observations on the same unit, identification of the education effect is obtained by assuming that appropriately transformed data will produce an observation with a transformed wage equation error which is orthogonal to the transformed education level. The policy shock approach, on the other hand, imposes orthogonality between the "instrument" or policy shock and the wage equation error. That is, it is explicitly assumed that the policy shock affects the individual's decision to invest in education but does not directly affect the individual's wage. Using different types of policy shocks on the same data may produce different estimates of the effect of education as they may identify the effect from different parts of the population (see Imbens and Angrist 1994).

Although estimates based on the conditional first moments possess desirable statistical features when the moment conditions are both "valid" and informative, there is concern that in many empirical examples there are no available instruments. There is also concern that inference based on such estimation is not reliable when the moments are "uninformative" due to problems related to the use of weak instruments (see, for example, Staiger and Stock 1997). Accordingly, we adopt an approach which bypasses the need to find instruments. Instead, we employ a methodology which combines restrictions on the conditional second moment with presence of heteroskedasticity in the model as the basis for identification.

We employ such an approach, described in detail below, to estimate the returns to schooling for a sample of Australian workers. Previous research on the returns to schooling in Australia has indicated that the individual's background features often directly affect earnings making it difficult to find appropriate exclusion restrictions to act as instruments. Also, there have been few relevant policy shocks which might be exploited to generate variation in educational attainment and there exist limited data sets which allow one to control for unit specific endogeneity. Moreover, the Australian tertiary educational system is currently the focus of an ongoing discussion about how much of the financial burden of higher education should be borne by the student. Clearly a reliable estimate of the return to education is crucial to this debate. The next section describes the estimation procedure, while Section III describes the data and estimation results. Section IV provides some concluding comments. Finally, note that while this paper focuses on the Australian evidence for the returns to education the identification approach we outline is appropriate to other settings and other economic problems.
II. Identification and Estimation Strategies

A. Model and Identification

Consider the following simultaneous (triangular) model for wages and education:

\[(1)\quad W_i = X_i\beta_0 + \beta_1 E_i + u_i, \quad i = 1..N\]
\[(2)\quad E_i = X_i\delta_0 + v_i\]

where the \(X\)-variables are uncorrelated with the errors \(u\) and \(v\) in both the wage (\(W\)) and education (\(E\)) equations. Note that the same \(X\)'s are allowed to appear in both equations. We assume that the \(u\) and \(v\) are correlated and thus OLS on the wage equation does not produce consistent estimates of the parameter \(\beta_1\).

To consistently estimate \(\beta_1\) a control procedure, which is equivalent to IV, is frequently employed where an estimate of \(v\) is incorporated into the wage equation with a constant impact. To explain this procedure and our subsequent modification, for notational convenience, we will not distinguish \(v\) from its estimate (residual) in what follows. Define:

\[\epsilon = u - \lambda_0 v, \quad \lambda_0 = [\text{cov}(u, v)/\sigma_v^2] .\]

The parameter \(\lambda_0\) is the population value or limiting value in large samples of the OLS coefficient obtained by regressing \(u\) on \(v\). By construction, once the impact of \(v\) is removed from \(u\), the resulting \(\epsilon\) component is uncorrelated with \(v\). This observation motivates estimating the controlled regression:

\[W_i = X_i\beta_0 + \beta_1 E_i + \lambda_0 v + \epsilon_i\]

where we control for endogeneity of \(E_i\) by including \(v\) in the wage equation. Without exclusion restrictions, however, the variables on the righthand-side are linearly dependent and the matrix of these variables \(D = [X, E, v]\) does not have full column rank. Accordingly, it is not possible to identify the wage equation parameters.

Note that we would be able to identify the coefficients of interest if the impact of the control was not constant (that is, dependent on \(X\)) and if it were possible to estimate this impact. To explore such a situation suppose that the joint distribution of \(u\) and \(v\) depends on \(X\). In particular, let the conditional variances of the errors depend on \(X\). Denoting \(S_v^2(X)\) as the conditional variance function for \(v\), define:

\[\epsilon \equiv u - A(X)v, \quad A(X) = [\text{cov}(u, v|X)/S_v^2(X)].\]

For each value of \(X\), \(\epsilon\) is again defined by removing the impact of \(v\) on \(u\). Therefore, conditioned on \(X\), \(\epsilon\) is uncorrelated with \(v\) by construction. Similar to the discussion above, we will refer to \(A(X)\) as the impact of the control. However, unlike the case above, this impact may now depend on \(X\) and therefore may not be constant.

1. In practice the residual is used in place of \(v\).
That is, the matrix $D_1 = [X,E,A(X)v]$ is of full rank. Thus, provided it is possible to consistently estimate the variable impact $A(X)$, the model can be identified without exclusion restrictions.

Below we will discuss several error structures where the control has this variable impact property (VIP) and for which Klein and Vella (Forthcoming) show that it is possible to consistently estimate $A(X)$. We first note that the VIP does not necessarily hold even when the joint distribution of $u$ and $v$ depends on $X$. For example, with $e$ as a random error component consider the additive error structure below with the conditional variances of $u$ and $v$ depending on $X$:

$$u = \lambda_0 v + \epsilon, \quad E(v|X) = E(\epsilon|X) = E(\nu\epsilon|X) = 0.$$  

In this model, $A(X)$ is the constant $\lambda_0$ even though both conditional variance functions may depend on $X$.

We now turn to several error structures under which the VIP holds and it is possible to consistently estimate $A(X)$. Denote $S_u(X)$ and $S_v(X)$ as the conditional variance functions for $u$ and $v$ respectively. Then with:

$$u = S_u(X)u^* \text{ and } v = S_v(X)v^*$$

consider the following additive structure for the unscaled errors.

$$u^* = \lambda_0 v^* + \epsilon^*, \quad E(u^*|X) = E(v^*|X) = 0; \text{ cov}(v^*,\epsilon^*|X) = 0.$$  

For this case, it follows that:

$$A(X) = \alpha_0 S_u(X)/S_v(X),$$

a variable impact which, as discussed below, can be consistently estimated. Alternatively, with $\epsilon_1$ and $\epsilon_2$ being mean-zero error components that are independent of $X$, consider the multiplicative error structure:

$$u = \alpha_1(X)w^*\epsilon_1; \quad v = \alpha_2(X)w^*\epsilon_2,$$

where $\epsilon_1$ and $\epsilon_2$ are independent of the common error component, $w^*$. The conditional second moment for the common error component, $w^*$, may or may not depend on $X$. In this case, with $\rho_0$ as the correlation between $\epsilon_1$ and $\epsilon_2$:

$$A(X) = [\text{cov}(u,v|X)/S_v^2(X)] = \rho_0[S_u(X)/S_v(X)],$$

a form identical to that in the previous example.

For both of the above error structures, the conditional correlation:

$$\rho_0 = \text{cov}(u,v|X = x)/[S_u(X)S_v(X)]$$

is constant. Whenever this condition holds, it follows in general that:
\[ A(X) = \left[ \text{cov}(u, v|X)/S_v^2(X) \right] = \rho_0[S_u(X)/S_v(X)]. \]

Klein and Vella (Forthcoming.) show that under the above constant correlation condition (CCC) it is possible to estimate \( A(X) \) consistently in a semiparametric formulation without specifying functional forms for the conditional variance functions. Provided that the ratio of these functions depends on \( X \), a condition typically satisfied when heteroskedasticity is present and conditional variance functions are not proportional to each other (for all \( X \) values), then the model is identified without exclusion restrictions.

B. Discussion

Before briefly describing the nature of the estimation method it is instructive to interpret the error structures we allow for in the context of the current application. To do so we consider the two major features of the error distributions that our approach requires. The first is the presence of heteroskedasticity, captured by the \( S_u \) and \( S_v \) functions respectively, and the second is the manner in which the error terms, reflecting the heterogeneity, are related across equations.

The presence of heteroskedasticity is usually an empirical question but it seems likely to arise in the setting we consider here. For example, in explaining the variation in the level of educational attainment across individuals a number of variables, such as those capturing regional factors and school types, are likely to encompass several underlying effects. For example, with regional factors it is likely that the location of the schools within a geographical region will not only affect the mean level of educational attainment but also the variance of the educational attainment in the region. For example, if the distance to the nearest school is the determining factor in an individual’s schooling attainment, as is discussed in Card (1995), consider the following two situations. In Region A all individuals are equidistant from their nearest school while in Region B some individuals are very close and others are very distant. In this scenario the allocation of schools within each region could be chosen such that it is not clear which region would have the highest average education. However, clearly the variance in schooling attainment in region A will be less than that in region B. Thus, one would expect a relationship between the variance in the schooling equation error and the regional variables. Heteroskedasticity will also arise if there are heterogeneous effects from particular variables or there are varying parameters in the models. In the case of heterogeneous effects consider the effect of attending a private school on educational attainment. Clearly there are likely to be some private schools that have very large effects while there are others that have relatively smaller effects. As the coefficient on the private school variable will capture some weighted average of these effects it is also likely to introduce heteroskedasticity. It is also possible that random parameters for the variables explaining the conditional mean of educational attainment will introduce heteroskedasticity which is related to the explanatory variables. That is, if the parameters \( \delta \) in Equation 2 are functions of the \( X \)’s.

Similar logic applies to the presence of heteroskedasticity in the wage equation. That is, in some instances the effect of the variable on wages may vary across individual depending on their characteristics. Moreover a heterogeneous effect for some
variables may also arise in the wage equation. Furthermore, many individual characteristics may directly influence the variance of the error. For example, attendance at certain types of schools could either decrease or increase the variance of the wage error by not only ensuring a minimum level of quality of education for the less talented individual but by also assisting the more talented individuals to do well in the labor market. Also, as individuals obtain greater work experience it is likely that some will do better than others in terms of wage growth. This will introduce heteroskedasticity as a function of job tenure.

We adopt a relatively agnostic approach to what is generating the heteroskedasticity as our approach does not require that we take a stance on its form. However, while we do not impose any restrictions on the form of the $S$ functions we assume below, for the sake of making estimation feasible, that the heteroskedasticity is a function of a single index in each equation. While this does restrict the classes of heteroskedasticity that we can incorporate this restriction is somewhat offset by our ability to estimate the heteroskedastic functions nonparametrically. Moreover, in any application the variables contained in the respective indices can be expanded to include functions of the explanatory variables if this was considered more reasonable. Thus while the $X$ variables may cause heteroskedasticity through a number of mechanisms we consider the single index approximation to be the leading case.

The endogeneity of education in wage equations is typically due to the presence of some unobservable factor(s) that influence both wages and education. While this is frequently interpreted as unobserved ability the empirical evidence is not always clearly in favor of such an interpretation and other possible candidates such as motivation, educational overachievement and measurement error are other candidate explanations.

Our remark that the evidence is not always clearly in favor of the unobserved ability interpretation is due to the common finding that accounting for the endogeneity of education typically leads to an increase in the schooling effect. To obtain such an increase in the schooling effect would generally require the correlation in the errors across equations, captured by the parameter $\rho$, to be negative while the unobserved ability story would require that $\rho$ be positive. That is, the unobservables increase both schooling and wages. We discuss this further below.

For the sake of interpretation we retain the unobserved ability interpretation. In this case the error process described in Equation 3 has the following interpretation. The error $v^*$ is the level of unobserved ability and its impact on the observed level of education depends on the individual's environment as captured by the $X$'s. The error term $u^*$ is a linear function of this endowment of unobserved ability and it subsequent impact on wages is scaled up by the individual's $X$'s. In this context the estimate of $\rho$ captures how unobserved ability affects both wages and education and one would suspect that this, in general, should be positive.

A second interesting case is that shown in Equation 4. Continue to assume that wages and education each depend on some unobserved factor $a^*$, which may reflect the individual's unobserved ability. Furthermore, assume that the impact of this

2. Klein and Vella (Forthcoming) provide a proof of identification in the absence of index restrictions in the heteroskedastic functions. However, estimating the models without imposing such restrictions in the $S_a$ and $S_v$ functions is likely to be very difficult when the $X$ vector is of large dimension.
factor is not constant in at least one of the equations. Let the impact of $a^*$ differ in these two equations and let it consist of a potentially predictable or estimable component that depends on $X$ and a random component that cannot be estimated. Denote $a_1(X)$ and $a_2(X)$ as the predictable impacts for wage and education equations respectively and let $\epsilon_1$ and $\epsilon_2$ be the corresponding unpredictable components which capture other possible influences. Under these conditions, the unobserved factor enters wage and education equations as:

$$W : u = a_1(X)a^*\epsilon_1$$
$$E : v = a_2(X)a^*\epsilon_2.$$  

With the components satisfying the conditions above, the control has a variable impact that depends on the ratio of conditional variance functions. In this case, most importantly, the coefficient $\rho$ captures the correlation between the $\epsilon_1$ and $\epsilon_2$ components and their impact on education and wages. Thus in this case the unobserved ability increases both wages and schooling but its interaction with the $\epsilon$'s may induce a negative correlation across equations. For example, following the interpretation of Vella and Gregory (1996) assume that $\epsilon_1$ captures some measure of "overachievement" in educational attainment. The value of $\epsilon_2$ would then reflect how this "overachievement" is valued in the labor market. In this case a positive $\rho$ would mean the overachievement augmented the wage while a negative value, as found in Vella and Gregory (1996), would mean that it decreased the wage. Note, however, that a negative estimate of $\rho$ would not necessarily imply that wages decrease as the level of $\epsilon_2$ increases. This is because an increase in $\epsilon_2$ is simultaneously associated with a higher wage resulting from the rise in education.

Our estimation procedure makes no restrictions on the data generating process underlying the data and we present the two cases above only to show that our model encompasses special cases of interest. Moreover, ultimately the sign of $\rho$ is a function of the data and its interpretation is seen through the eyes of the practitioner. In our empirical work that follows we offer one interpretation of our estimate of $\rho$ but this does not exclude others from adopting an alternative interpretation. It is interesting to note, however, that the above discussion illustrates that it is difficult to interpret $\rho$ unless one imposes a very strict, and perhaps unrealistic, structure on the disturbances. Moreover, this consideration is not unique to our approach but is always relevant when economic inference is based on the correlation of residuals across equations.

**C. Estimation**

Although it is possible to estimate $A(X)$ nonparametrically, it is more practical to impose an index structure in order to obtain reasonable estimates at moderate sample sizes. Namely, assume that conditional variance functions both depend on linear indices:

$$S^2_v(X) = S^2_v(X\alpha_0); \quad S^2_u(X) = S^2_u(X\gamma_0).$$

where $\alpha_0$ and $\gamma_0$ are unknown parameters. With $\hat{E}$ as a nonparametric conditional expectations estimator, we employ Semiparametric Least-Squares (see Ichimura
1993), to estimate the parameters of the \( v \)-index. Namely, subject to the usual normalization on \( \alpha_0 \)\(^3\)

\[
\hat{\alpha} = \arg \min_{\alpha} \sum_{i=1}^{N} \left( \hat{v}_i^2 - \hat{E}(\hat{v}_i^2 | X \alpha) \right)^2 / N,
\]

where \( \hat{v}_i^2 \) is a consistently estimated squared residual. It then follows that the conditional variance function for \( v \) is given as:

\[
\hat{S}_v^2(X) = \hat{E}(\hat{v}_i^2 | X \hat{\alpha}).
\]

The other conditional variance function must be estimated simultaneously along with other parameters of interest. To outline the nature of the estimation method, define:

\[
u_i(\beta) = W_i - X_i \beta_0 - \beta_1 E_i
\]

where the \( \beta \)'s are arbitrary parameter values. Notice that at the true parameter values that \( u_i(\beta_0) = u_i \), the wage error. Next, define a "variance-type" function:

\[
S_{\mu}^2(\beta, \gamma) = E[u_i^2(\beta)|X_i \gamma].
\]

Notice that at the true parameter values:

\[
S_{\mu}^2(\beta_0, \gamma_0) = S_u^2(X_i);
\]

the true conditional variance function for \( u_i \). Replacing the true expectation \( E \) above with the nonparametric estimator \( \hat{E} \), we obtain the feasible estimator:

\[
\hat{S}_{\mu}^2(\beta, \gamma) = \hat{E}[u_i^2(\beta)|X_i \gamma].
\]

We then consider the following controlled nonlinear model:

\[
W_i = X_i \beta_o + E_i \beta_{1o} + \hat{A}_i(\beta, \gamma) v_i + error.
\]

Parameter estimates are then "essentially" obtained by selecting \( \gamma \) and \( \beta \) parameters to minimize the sum of squared residuals.\(^4\) Klein and Vella (Forthcoming) establish that the resulting estimator is consistent and asymptotically distributed as normal at the usual \( \sqrt{N} \) parametric rate.

---

3. With the first variable in \( X \) being a continuous variable and with \( X \) not containing a column of ones: \( \alpha = [1; \alpha] \). Subject to this normalization (much like setting the variance of the error in a probit model to 1), the \( \alpha_2 \) are identified. In a probit model, we are able to identify probability functions under the normalization that the variance is 1. Similarly, here conditional variance functions are identified once we know the normalized parameters above.

4. Here, "essentially" means that, as shown in Klein and Vella (Forthcoming), several different nonparametric expectations operators are required to obtain identification. One expectation is based on the \( u \) index while the other is based on both \( u \) and \( v \) indices. While the reader is referred to Klein and Vella (Forthcoming) for a detailed discussion of the identification argument, the appendix provided here gives a detailed description of the estimation procedure.
The general approach employed here is related to other estimators used in this context. The rank order estimator of Vella and Verbeek (1997), applied to the returns to education in Rummery, Vella, and Verbeek (1999), is also identified via heteroskedasticity. The Vella and Verbeek (1997) estimator requires that rank order of the individual's residual, in a given subset of the data, is relevant rather than the value of the residual itself. They then construct an instrumental-variable procedure based on this premise. This is a special case of the approach employed here. However, while Vella and Verbeek (1997) assume that $S_u$ and $S_v$ are uncorrelated, this is not imposed here. Another related paper is Hogan and Rigobon (2004), who form an alternative structure for $A(X)$ in that they focus on conditional covariances. That is, they assume that some variable is related to the variances of the education equation error but does not directly determine wages. They also do not estimate the model in the control function manner but follow the procedure outlined in Rigobon (2003). Note that the underlying identification strategy they employ is closely related to the Vella and Verbeek (1997) approach.

III. Data and Model Specification

To estimate the impact of schooling on earnings, we employ the 2001 wave of “The Household, Income and Labour Dynamics in Australia (HILDA) Survey.” These data contain labor market and background information on a sample of Australians. We examine the determinants of wages for a sample of 5070 working individuals living in the five most populous states in Australia.\(^5\) We focus on the wage determination process conditional on working and do not address any sample selectivity issues induced by the working decision.

Estimating the impact of education in the Australian context is an interesting problem. Arguing that the high returns to education merited the additional investment, the Australian Federal Government actively encouraged individuals to seek additional education (Vella and Gregory 1996). There is also a very active ongoing debate regarding who should bear the cost of such investment. That is, there are proposals to shift an increasing share of the cost of tertiary education onto the students undertaking the investment.

An important contribution to this debate would be a reliable estimate of the return to schooling. Previous papers have supported the conjecture that education is endogenous to wages thereby invalidating the evidence from studies which do not account for the endogeneity. However, obtaining estimates that do account for the endogeneity is complicated by the lack of instruments in the Australian context. For example, Vella and Gregory (1996) provide evidence that the individual's background characteristics directly influenced wages, making it difficult to assign background characteristics the role of instruments. Leigh and Ryan (2005) employ a strategy similar to Angrist and Krueger (1991) but find that the instruments do not provide an estimate of the return to education with tight bounds in the Australian context. Miller, Mulvey and Martin (1995) employ a sample of twins to estimate the returns to

---

\(^5\) We exclude observations from the less populous states and regions to avoid estimation issues related to small cell sizes.
schooling, but, as the data do not contain a direct measure of wages, it is necessary to impute them. Given this evidence we employ the estimation strategy outlined in the previous section where we employ all the variables in the schooling equation as regressors in the wage regression and identify the schooling effect via heteroskedasticity and restrictions on the conditional second moments.

The variables used in the estimation of the model, and their summary statistics, are reported in Table 1. The estimates for the education and wage equations are reported in Tables 2 and 3 respectively, noting that the dependent variables are the number of years of schooling \((school)\) and the log of the hourly wage rate \((wage)\). The specifications contain the typical variables used in the estimation of schooling and wage equations, and capture the individual's background characteristics and some features of the schooling type and location. We also include two variables that reflect the amount of time the individual has been in Australia. These are \textit{Born in Australia}, which denotes the individual was born in Australia, and \textit{Years in Australia} which captures the number of years the immigrants in the sample have been in Australia. We also employ additional, work-related variables in addition to marital status, in the wage equation which do not appear in the education equation. These do not identify the model as conventional IV procedures require a variable(s) in the education equation which does not appear in the wage equation. We employ the same exogenous variables for each of these indices for conditional variances that appear in the corresponding conditional means. Note that the semiparametric nature of the estimation procedure requires the existence of continuously distributed variables. For the schooling education we require one such variable and the variable we employ is the individual's age. For the wage equation we require two variables given the nature of the conditioning and we employ the individual's level of job tenure in addition to the individual's age.

In Table 2 we report the estimates for the schooling equation. Note that for the schooling equation we highlight again that we use the same explanatory variables for both the conditional mean and for the heteroskedasticity index. The absolute values of the \(t\) statistics for the estimates are reported in parentheses. As expected, a number of the individual's background characteristics have a statistically significant impact on the level of acquired schooling. Australian-born individuals acquire approximately one year of education less than relatively recently arrived foreign born individuals but this difference decreases as the number of years in Australia increases. This latter effect probably reflects when the individual arrived in Australia. Having both parents present when the individual was aged 14 years increases educational attainment by about half of one year. While this may reflect parental guidance or household stability, it is also likely to capture income effects. The presence of siblings also decreases educational attainment. Attendance at Catholic or other types of private schools has a very large and statistically significant positive effect on the level of education obtained, recalling that the excluded group is those attending a government-financed school. The \textit{Private School} coefficient indicates that attendance at such a school increases educational attainment by 0.87 years as opposed to attendance at a government school. The regional variables indicate some differences across States. Finally, males acquire 0.3 years of schooling less than females.

As the coefficients in the index for the heteroskedastic function have no immediate interpretation we do not report them here. We note, however, that several of them, including the continuous age variables were statistically significant. Moreover, the
### Table 1

**Summary Statistics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hourly wage in Australian dollars</td>
<td>20.538</td>
<td>10.823</td>
</tr>
<tr>
<td>Years of education (school)</td>
<td>12.470</td>
<td>2.222</td>
</tr>
<tr>
<td>Born in Australia</td>
<td>0.778</td>
<td></td>
</tr>
<tr>
<td>Both parents present at age 14</td>
<td>0.829</td>
<td></td>
</tr>
<tr>
<td>Number of siblings</td>
<td>2.727</td>
<td>1.977</td>
</tr>
<tr>
<td>Father unemployed when individual was young</td>
<td>0.095</td>
<td></td>
</tr>
<tr>
<td>Mother employed when individual was young</td>
<td>0.520</td>
<td></td>
</tr>
<tr>
<td>Individual attended private school</td>
<td>0.255</td>
<td></td>
</tr>
<tr>
<td>School located in New South Wales</td>
<td>0.312</td>
<td></td>
</tr>
<tr>
<td>School located in Victoria</td>
<td>0.257</td>
<td></td>
</tr>
<tr>
<td>School located in Queensland</td>
<td>0.221</td>
<td></td>
</tr>
<tr>
<td>School located in South Australia</td>
<td>0.094</td>
<td></td>
</tr>
<tr>
<td>School located in Western Australia</td>
<td>0.104</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>0.544</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.536</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>38.488</td>
<td>11.522</td>
</tr>
<tr>
<td>Years in Australia</td>
<td>4.503</td>
<td>10.573</td>
</tr>
<tr>
<td>Job tenure</td>
<td>6.488</td>
<td>7.604</td>
</tr>
</tbody>
</table>

$NR^2$ from the regression of the residuals squared on the all of the exogenous explanatory variables employed in the schooling equation produced a value of 263.39. This test value strongly rejects the null hypothesis of homoskedastic errors. Moreover, several of the groups of variables, as well as some individually, appeared to be a source of heteroskedasticity.

In Table 3 we report the estimates for the wage equation. Once again we use all the variables, with the exception of the schooling variable, that appear in the conditional mean for wages in the index underlying the heteroskedasticity in the wage equation. The absolute values of the t statistics are reported in parentheses. First consider the OLS equation as a number of its features are worth noting. There is some evidence that the background variables have direct influence on the wage level when the endogeneity of education is not taken into account. The number of siblings, for example, appears to directly decrease the wage. This may be explained by the quality of education one obtains in the presence of several siblings if there are tradeoffs with quality as well as quantity as indicated in Table 2. There is also a positive relationship between the level of earnings and whether or not the individual’s mother worked but the effect is only relatively weak in terms of statistical significance. Attendance at a private school increases wages by 5.5 percent and this effect would be statistically significant in the absence of the endogeneity problem. There is also evidence of a small marriage premium and gender differential of 10 percent in favor of males.
Table 2

*Education Equation*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>10.254**</td>
<td>Victoria</td>
<td>0.075</td>
</tr>
<tr>
<td>Age</td>
<td>0.192**</td>
<td>Queensland</td>
<td>-0.305**</td>
</tr>
<tr>
<td>Age²</td>
<td>-0.003**</td>
<td>South Australia</td>
<td>0.079</td>
</tr>
<tr>
<td>Both parents present</td>
<td>0.539**</td>
<td>Western Australia</td>
<td>-0.082</td>
</tr>
<tr>
<td>Number of siblings</td>
<td>-0.176**</td>
<td>Born in Australia</td>
<td>-1.018**</td>
</tr>
<tr>
<td>Father unemployed</td>
<td>-0.158</td>
<td>Male</td>
<td>-0.313**</td>
</tr>
<tr>
<td>Mother employed</td>
<td>0.116*</td>
<td>Years in Australia</td>
<td>-0.029**</td>
</tr>
<tr>
<td>Private school</td>
<td>0.871**</td>
<td></td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: "*" and "**" denote significance at the 5 and 10 percent levels respectively.

The regional variables are statistically significant but this most likely reflects the higher cost of living in the control group NSW noting that most people are likely to be living in the state in which they attended school. Finally, the point estimate for the education effect is 6 percent.

Now focus on the estimates that are adjusted for the endogeneity of education via the inclusion of our proposed control function. These are reported under the heading CF alongside the OLS estimates in Table 3. Once again, the absolute values of the t statistics appear in parentheses, noting that the manner in which they are estimated accounts for the nature of the estimation process. Before we focus on the coefficient of primary interest, it is valuable to note that while the estimates across the two columns for the exogenous variables are generally quite similar there are some important differences. More importantly these differences have implications for what may or may not be reasonable instruments. More explicitly, this column indicates that, on the basis of their lack of statistical significance, two possible instruments may be the number of siblings or school type although each of these was statistically significant in the OLS estimation. The key feature of this column, however, is the estimate of the education coefficient. While the OLS estimate was 6 percent we see that the CF estimate is 10 percent. Moreover, while there is some loss in statistical significance, in comparison to the OLS estimate, the coefficient is statistically significant at levels far below 5 percent. This estimate indicates that the return to education obtained through our procedure is notably higher than that indicated through OLS estimation.

The increase in the return to schooling is due to the control function having an estimated coefficient of -0.203. This negative coefficient indicates that the unobservables correlated with wages are negatively correlated with education. This finding is consistent with the results of Vella and Gregory (1996), who interpreted such a result...
### Table 3
**Wage Equation**

<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS</th>
<th>CF</th>
<th>Variable</th>
<th>OLS</th>
<th>CF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.947** (12.387)</td>
<td>0.573** (2.615)</td>
<td>South Australia</td>
<td>-0.121** (5.691)</td>
<td>0.130** (5.838)</td>
</tr>
<tr>
<td>Age</td>
<td>0.051** (14.591)</td>
<td>0.042** (7.909)</td>
<td>Western Australia</td>
<td>-0.052** (2.577)</td>
<td>-0.064** (2.955)</td>
</tr>
<tr>
<td>Age²</td>
<td>-0.0005** (13.176)</td>
<td>-0.0005** (6.425)</td>
<td>Born in Australia</td>
<td>0.089** (3.734)</td>
<td>0.121** (3.527)</td>
</tr>
<tr>
<td>Both parents present</td>
<td>-0.003 (0.206)</td>
<td>-0.026 (1.247)</td>
<td>Married</td>
<td>0.036** (2.781)</td>
<td>0.040** (2.955)</td>
</tr>
<tr>
<td>Number of siblings</td>
<td>-0.006** (2.235)</td>
<td>-0.001 (0.142)</td>
<td>Male</td>
<td>0.097** (8.414)</td>
<td>0.121** (8.264)</td>
</tr>
<tr>
<td>Father unemployed</td>
<td>0.002 (0.123)</td>
<td>0.006 (0.292)</td>
<td>Years in Australia</td>
<td>0.004** (4.955)</td>
<td>0.005** (3.606)</td>
</tr>
<tr>
<td>Mother employed</td>
<td>0.017 (1.447)</td>
<td>0.017 (1.327)</td>
<td>Tenure</td>
<td>0.012** (5.605)</td>
<td>0.011** (3.803)</td>
</tr>
<tr>
<td>Private school</td>
<td>0.055** (4.075)</td>
<td>0.018 (0.798)</td>
<td>Tenure²</td>
<td>-0.0001** (2.145)</td>
<td>-0.0001 (1.163)</td>
</tr>
<tr>
<td>Victoria</td>
<td>-0.034 (2.257)</td>
<td>-0.042 (2.560)</td>
<td>School</td>
<td>0.060** (21.757)</td>
<td>0.100** (5.260)</td>
</tr>
<tr>
<td>Queensland</td>
<td>-0.081 (5.130)</td>
<td>-0.073 (4.294)</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R²</td>
<td>0.223</td>
<td></td>
</tr>
</tbody>
</table>

Note: ** and * denote significance at the 5 and 10 percent levels respectively.
as a "penalty" to educational overachievement. That is, the unobserved factors that increased an individual's education level above what was predicted by his or her background characteristics have a negative impact on the wage level. Thus individuals who obtain a level of education above what is predicted by the model received less for their incremental increase in education than those who were predicted to obtain that level.

While the return to education is somewhat offset by this "penalty" operating through the control function, the large increase in the point estimate indicates that the return to education is still high and markedly higher than the OLS estimate. This finding is consistent with the general finding that accounting for the endogeneity of education leads to an increase in the estimated return to education in general and also, more specifically, in the Australian context. Vella and Gregory (1996), for example, find that accounting for the endogeneity of education greatly increased the estimated return for Australian youth.

To further interpret our results, it appears that the unobserved education residual does not capture only unobserved ability. If it did the negative coefficient on \( p \) would suggest that the return to unobserved ability is negative. In contrast, and following our earlier discussion, we interpret the errors as capturing both unobserved ability and some other factors. Thus, as discussed above, it is possible that \( p \) is capturing the correlation between these other factors across the two equations. We interpret these other factors to capture considerations that are leading the individual to attain a higher than expected, on the basis of his/her characteristics, level of education. While these factors lead to higher wages through the high returns to education their direct effect on wages is negative.

As with the schooling equation there is some indication of heteroskedasticity in the wage equation. The \( NR^2 \) for the regression of the residuals squared on the exogenous variables produced a value of 110.78, which indicates there is less evidence of heteroskedasticity in this equation than in the schooling equation. However, while there is less heteroskedasticity in the wage equation, this does not affect the validity of the estimator as heteroskedasticity in only one of the equations is sufficient.

Before proceeding to considering IV estimates of the return to schooling with these data, it is also worth highlighting that the coefficient \( p \) is statistically different from zero at the 5 percent level. This indicates that education is endogenous to wages. This is an important result, as it implies that the identifying moment is sufficiently informative to identify that education is endogenous to earnings.

The empirical investigation clearly highlights the issue associated with this uncertain choice of instruments. Many of the background variables have statistical significance levels which make it unclear whether they can be employed as instruments. That is, many of them are marginal in the CF estimation in terms of statistical significance. Consequently, neglecting pretest issues, it is possible that they can be employed as instruments. Accordingly, employing the same data we reestimated the model where we used the same first step, shown in Table 2, and in the second we employed the predicted value as an instrument for education.

---

6. The negative coefficient is also consistent with measurement error in the schooling variable although there does not appear to be any reason to suspect this is an issue here.
Given the estimates in Table 3 from the control function estimator it appears that the background variables plus school type are potential instruments. This is interesting since the OLS estimates in Column 1 of this table indicated that the number of siblings and school type were not valid instruments as they had a direct effect on wages. Accordingly, we first employed Both Parents Present, Father Unemployed, and Mother Employed as the only instruments. Doing so provided an estimate of the schooling effect of 0.063 with a t-statistic of 2.42. Not surprisingly, given the small increase in the education coefficient in comparison to the OLS estimate, the t-statistic on the residual is -0.12 indicating that one could not reject exogeneity of schooling.\(^7\)

The control function estimates suggest that both # of Siblings and Private School are both potential instruments. Accordingly, we excluded each separately from the wage equation to estimate the schooling effect. Including # of Siblings as an instrument with the other background variables while including the school type variable in the wage equation produced an estimate of 0.090 with a t-statistic of 6.76 for the schooling variable. The coefficient on the residual is -2.19 which rejects exogeneity. Including Private School as an instrument, while keeping # of siblings in the wage equation, produced an estimate of 0.102 (t = 8.48) and a t-statistic on the residual of -3.49. Finally using all the background variables plus # of siblings and Private School produced an estimate of 0.101 (t = 10.46) and a t-statistic on the residual of -4.14.

It is interesting to note that the coefficient on this residual is also negative and, similar to the estimates from our estimation procedure; this is inconsistent with the disturbances capturing only unobserved ability. Thus, the IV estimates support the interpretation of our control function estimates that the disturbances in these equations do not solely reflect unobserved ability.

Continuing to neglect pretest issues, these IV estimates provide support for our approach. First, the OLS estimates illustrate the inappropriateness of trying to identify instruments from a regression that does not account for the endogeneity. Second, while several of the background variables appeared to be valid instruments, they were unable to identify the endogeneity of the schooling variable. Third, the estimates based on schooling type and number of siblings were able to identify the presence of endogeneity. However, without our approach being first employed, one would not have known if these were valid instruments. Finally, the resulting IV estimate provided an estimate of the return to schooling, which is almost identical to the estimate based on our control function procedure. As this IV estimate is not based on the CCC assumption, this appears to provide strong support for our approach in this setting. Overall, the evidence is quite suggestive that the return to schooling for the individuals in this data set is underestimated in OLS estimation and accounting for the bias in OLS estimation leads to an estimate of around 10 percent.

**IV. Conclusions**

This paper provides an alternative approach to identifying and estimating the returns to education in a model where the level of education is

---

7. The estimated education coefficient was obtained via IV, while the t-statistic for the residuals was obtained via the control function version of IV.
endogenous to wages. The identification strategy is based on combining the presence of heteroskedasticity in the model with the assumption that the correlation between the errors, conditional on the exogenous variables, is constant. For a sample of Australian workers we find a large increase in the return to schooling in comparison to what is found when the endogeneity is not accounted for. More explicitly, we obtain an estimate of 10 percent in contrast to the OLS estimate of 6 percent. We also find that using our approach we are able to identify variables that would otherwise have been excluded as instruments. Using these variables as instruments provides an estimate of the return to schooling that is almost identical to our estimate thereby providing support for our identifying strategy.

Finally we note that the estimation strategy we employ here is complicated by the semiparametric treatment of the heteroskedastic functions. We adopt such an approach to avoid assumptions regarding these functions. However, in practice one may find it attractive to make parametric assumptions regarding the form of the heteroskedasticity. In these instances one can still rely on the same identification strategy, but in implementing the procedure there is no need to augment the criterion function in the second step with the component that conditions on both indices simultaneously. This would greatly reduce the required computation. It also eliminates the need for continuously distributed exogenous variables in the indices. Given the difficulty in obtaining instruments in many empirical investigations, we feel that our approach not only allows the estimation of the effect of education on wages in the absence of suitable instruments, but also provides a framework for studying many other economic problems of interest where the issue of endogeneity arises.

### Appendix

This appendix is based on the discussion in Klein and Vella (Forthcoming), hereafter KV, and outlines the two step estimation procedure.

First regress $E_i$ on $X_i$ to obtain $\delta$ and define the reduced form residuals as:

$$\hat{v}_i = E_i - X_i \hat{\delta}.$$  

To estimate the conditional variance function, $S^2_{\hat{v}}$, with a single index structure write:

$$S^2_{\hat{v}} = E[\hat{v}_i^2 | X_i] = E[\hat{v}_i^2 | I_{vi}(\alpha_0)],$$

where $I_{vi}(\alpha_0) = X_{1i} + X_{2i} \alpha_0$ where the $0$ subscript denotes the truth. We then estimate $\hat{\alpha}_0$ using semiparametric least squares with $\hat{v}_i^2$ as the dependent variable (see Ichimura 1993) as:

$$\alpha = \arg \min_{\alpha} \sum_{i=1}^N \hat{r}_i [\hat{v}_i^2 - E(\hat{v}_i^2 | I_{vi}(\alpha))]^2,$$

where $\hat{r}_i$ is a trimming function that restricts $X_i$ to a compact set depending on sample quantiles. The estimator for the conditional variance function is then given as:
where $\hat{E}$ is a nonparametric estimator for the indicated conditional expectation. Employing the above initial estimator $\hat{S}_{vi}$, we then repeat the above process in a GLS step. In estimating the wage equation parameters we begin by noting consistent residuals are not immediately available. Accordingly, the conditional variance function for this equation and the parameters of interest are estimated simultaneously. Once again we impose an index restriction but KV show that identification under such a restriction requires that we augment the criterion function which corresponds to that used for reduced form in such a way to ensure we are achieving the correct minimum. We provide the criterion function below although the reader is referred to KV for the argument underlying the identification strategy.

We construct the appropriate index for the wage equation heteroskedasticity as $l_{ui}(\gamma_o) = X_{i1} + X_{i2}\gamma_o$. Note that for generality we allow the same $X$'s to appear in both $l_u$ and $l_v$ although this is not necessary and not imposed in the empirical work. Given the nature of the approach adopted for the education equation a natural approach to obtaining the wage equation estimates would be to estimate $\pi = (\beta, \gamma, \rho)$ as the following:

$$\pi = \arg\min_{\pi} \hat{Q}_1(\pi).$$

where:

$$u_i(\beta) = (W_i - X_i\beta_0 - \beta_1E_i); \quad \hat{S}_{ui}(\beta, \gamma)^2 = \hat{E}[u_i^2(\beta)|l_{ui}(\gamma_o)].$$

and:

$$\hat{A}_i(\pi) = \rho[\hat{S}_{ui}^2/\hat{S}_{vi}];$$

$$\hat{Q}_1(\pi) = (1/N) \sum_{i=1}^{N} \hat{v}_i[W_i - X_i\beta_0 - \beta_1E_i - \hat{A}_i(\pi)(\pi)\hat{v}_i]^2.$$ 

However, KV show minimizing $\hat{Q}_1(\pi)$ may not ensure that the correct minimum is obtained. Namely, in minimizing the probability limit of $\hat{Q}_1(\pi)$, the set of potential minimizers is not sufficiently restricted to enable an identification argument. In particular, these restrictions do not insure that $u$'s conditional variance function satisfies a single index restriction. To ensure that the set of minimizers does satisfy an appropriate index restriction, KV modify the objective function.

---

8. While it is possible to avoid a GLS step, we have found in a Monte Carlo investigation of the procedure that the estimator for $S_{vi}$ based on a GLS estimate of the parameters improved and that there is a corresponding improvement in the estimates of the primary equation.
Let
\[ \hat{S}_{ui}^{*2}(\beta, \gamma) = \hat{E}(u_i^2(\beta)|I_{ui}(\gamma), \hat{v}_i) \]
\[ \bar{\lambda}_{i}(\pi) = \rho[S_{ui}^{*}/\hat{S}_{vi}] \]
\[ \hat{Q}_2(\pi) = (1/N) \sum_{i=1}^{N} \hat{\tau}_i[w_i - x_i\beta_0 - \beta_1 e_i - \bar{\lambda}_i^{*}(\pi)\hat{v}_i]^2. \]

Here, \( \hat{S}_{ui}^{*2} \) captures the conditional expectation of the variance function when one conditions on its own index plus the additional index characterizing the data generating process. The final, combined objective function is then given as:
\[ \hat{Q}(\pi) = \hat{Q}_1(\pi) + \hat{Q}_2(\pi). \]

Denote \( Q, Q_I, \) and \( Q_2 \) as the limiting values (uniform probability limits) for the above objective functions. KV show that \( TTO, \) the vector of true parameter values, is a minimizer not only for \( Q \) but also separately for \( Q_I \) and \( Q_2. \) It follows that any other potential minimizer, \( \pi^* \), must also minimize each of these component objective functions. Consequently, any potential minimizer must satisfy restrictions that are implied by minimizing each objective function separately. As a result, KV show that for any candidate minimizer, the conditional variance function for \( u \) must satisfy an index restriction, a result that makes it possible to prove identification.

References


