ABSTRACT
The aim of the project is to estimate the returns to schooling for a data set containing the earnings of 540 individuals, their years of schooling and a set of nine control variables. After motivating the initial model and preliminary data transformation the initial results of the first two-stage-least-squares (TSLS) regression are presented. Insignificant variables are then dropped sequentially with the use of a two-sided p-value of 5 percent. Then a second TSLS together with a Hausman and Weak Instrument (WI) Test are presented and the Shea partial R2 statistic is computed for the relevance of the instrumental variables (IVs). The final model excludes the IV siblings due to its irrelevance. The model is tested for heteroskedasticity and the TSLS estimates are compared to the equivalent OLS regressions. Returns to schooling are estimated at 18.30 percent in the final model. However, this estimate has to be treated with caution.

Jasmin Droege
09/12/2015
# Table of Contents

- **Introduction**.................................................................................................................. 3
- Motivation for the Econometric Model .................................................................................. 3
- Empirical Results .................................................................................................................. 3
  - Final adjusted TSLS Model................................................................................................. 4
  - Comparison TSLS and OLS Results.................................................................................... 5
- Conclusion .............................................................................................................................. 5
- References ................................................................................................................................ 6
- Appendices ............................................................................................................................... 7
  - Summary Statistics ............................................................................................................. 7
  - Earnings vs Log Earnings .................................................................................................... 7
  - Q-Q Plot and Residual Plot TSLS 3 ................................................................................... 7
  - Variance-Covariance Matrix ............................................................................................. 8
  - TSLS and OLS Regression Results .................................................................................... 8
  - Shea Partial R-squared TSLS 2 .......................................................................................... 8
Introduction

The aim of the project is to estimate the returns to schooling. After motivating the model and presenting initial results, insignificant variables are dropped sequentially. The second TSLS model is tested with a Hausman and Weak Instrument Test. The Shea $R^2$ statistic suggests that the IV siblings is irrelevant and it is therefore excluded in TSLS 3. The final model is tested for heteroskedasticity and the estimates are compared to the equivalent OLS regressions.

Motivation for the Econometric Model

To estimate the returns to schooling log earnings $y_i$ is used as dependent variable and years of schooling $X_i$ as explanatory variable. In addition a set of nine potential control variables $W_i$ is available (appendix 1). As $y_i$ is positively skewed (appendix 2) and suffers from heteroskedasticity, its natural logarithm is deployed in order that the residuals are closer to being normally distributed and homoskedastic (Chiswick, 2003). The explanatory variable and the error $\varepsilon_i$ are anticipated to be influenced by the underlying talent of an individual leading to correlation of the error and the explanatory variable (Koop, 2008). This causes OLS to be inconsistent and therefore instruments $Z_i$ are deployed as a proxy for talent.

\[
\ln Y = \beta_0 + \beta_1 X_s + \beta_2 W_{wexp} + \beta_3 W_{female} + \beta_4 W_{ethblack} + \beta_5 W_{ethhisp} + \beta_6 W_{age} + \beta_7 W_{married} + \varepsilon
\]

Five exogenous regressors are included in the initial model. The variables gender, marital status and ethnicity are included to allow for demographic wage differences. Age and work experience are included due to accumulation of human capital. Years of schooling of father and mother as well as the number of siblings are chosen as IVs based the reasoning that children’s schooling is correlated with parents’ education and that siblings have identical abilities so that they can serve as a within-family estimator that will allow for an unbiased estimate (Card, 1999). The variance-covariance matrix (appendix 3) supports the assumption that the covariance of the endogenous explanatory variable and IVs is sufficiently strong.

Empirical Results

The initial regression model estimates returns to schooling at 16.74 percent\(^1\) (appendix 2). The Hausman test can only reject the null hypothesis that OLS estimates are consistent at a significance level of 10 percent due to a p-value of 0.0967. The Weak Instrument (WI) test confirms that the instrumental variables are not weak due to the first-stage F-statistic

\[^1 e^{0.154775} - 1 = 0.167395\]
\[ F(3, 530) = 46.52 > 10. \] However, the Sargan test rejects the null hypothesis that all instruments are valid at a significance level of 1 percent.

The variables Hispanic ethnicity, age and marital status turn out to be insignificant whereas Black ethnicity is significant at 10 percent level. Eliminating insignificant variables sequentially using a two-sided p-value of 0.05 results in a reduced regression of log income on schooling, work experience, gender and Black ethnicity. Through sequential elimination Black ethnicity becomes significant at 5 percent level and is therefore kept in the model. The return to schooling is estimated at 17.67 percent\(^2\). The Hausman test rejects the null hypothesis that OLS is consistent at a significance level of 5 percent. It can be concluded that OLS is inconsistent and, given this, TSLS is needed because the Gauss-Markov Theorem no longer holds under which OLS is BLUE (Koop, 2008). The Weak Instrument test confirms that the instruments are not weak but the Sargan test rejects the validity of the overidentifying restrictions again. However, this does not allow for the conclusion that the instruments are not able to identify the endogenous variable due to the nature of the test and the characteristics of the IVs (Parente and Silva, 2011). To investigate the relevance of the IVs the Shea partial \(R^2\) statistic is computed according to the simplification proposed by Godfrey (1999, appendix 6). From the table it can be inferred that both schooling of father and mother are relevant alone and in combination but that the IV siblings is irrelevant. Overall it can be noted that all instruments are rather weak.

\[
R^2 = \frac{\text{Var}(\beta_{\text{OLS}})}{\text{Var}(\beta_{\text{IV}})} = \frac{\sigma_{\text{IV}}^2}{\sigma_{\text{OLS}}^2} = \frac{\text{Var}(\beta_{\text{OLS}})}{\text{Var}(\beta_{\text{IV}})} \cdot \frac{1 - R^2_{\text{IV}}}{1 - R^2_{\text{OLS}}} \approx 0.21 \in [0, 1]
\]

**Final adjusted TSLS Model**

The third and final TSLS excludes the IV siblings on the basis of the Shea partial \(R^2\) statistic. For this the Hausman test again rejects the null hypothesis that OLS is consistent at a significance level of 5 percent and also the Weak Instrument test confirms that the instruments are not weak with a first-stage F-statistic of \(F(2, 534) = 69.78 > 10\). The third TSLS model is tested for heteroskedasticity with the use of a QQ-/residual plot (appendix 5) and the Pesaran-Taylor Test. For this the squared IV residuals \(\hat{\epsilon}_{\text{IV}}^2\) are regressed on the squared fitted values of the TSLS. The t-statistic of \(\hat{y}_{\text{TSLS}}^2\) is the appropriate test statistic \(HET_1 = 0.87 (0.38)\) and as is found to be insignificant, it can be concluded that the model does not suffer from heteroskedasticity.

\[ \hat{\theta}_{\text{IV}}^2 = \gamma_0 + \gamma_1 \hat{y}_{\text{TSLS}}^2 + \epsilon \]

\(^2e^{0.162691} - 1 = 0.176673\)
The plots reinforce the result and confirm that the residuals are normally distributed and that there is no clear pattern in their distribution.

**Comparison TSLS and OLS Results**

The return to schooling in the first equivalent OLS model is estimated at 13.34 percent$^3$. In this model also Hispanic ethnicity, age and marital status lack explanatory power. The second OLS model estimates the return to schooling at 13.57 percent$^4$ with all variables being significant at 1 percent level. In comparison the obtained OLS estimates are downward biased. This is at odds with the motivation for IVs in the project but may be explained by a measurement error and heterogeneity in the population that offset a modest upward ability bias (Card, 2001). On the other hand there is evidence that IV estimates may be more inconsistent than OLS estimates due to weak correlation of the IVs and the endogenous variable and potential finite sample bias due to a rather small sample (Bound, Jaeger and Baker, 1995).

**Conclusion**

The final model (TSLS 3) regresses log earnings on schooling, work experience and Black ethnicity. Thereby only schooling of father and mother are deployed as instrumental variables. The Hausman test suggests that OLS is inconsistent and that schooling is endogenous. Heteroskedasticity is not present in the model. Returns to schooling are estimated at 18.30 percent$^5$. This is significantly higher than the world average rate of return of 10.0 percent for 1970 to 2011 (Montenegro and Patrinos, 2013) which casts doubt on the validity of the obtained estimate. In the model females earn 25.60 percent$^6$ less than their counterparts all other things being equal. Likewise Black ethnicity reduces earnings by around 13.53 percent$^7$. This is strong evidence for gender and race discrimination due to significant unexplained group differences in earnings (Griliches, 1977). Besides, an additional year of work experience is estimated to increase earnings by 3.96 percent$^8$.

Overall the obtained estimates are not sufficiently robust due to their sensitivity to small changes in the model. The Sargan test, the Shea $R^2_S$ statistic and the Hausman test which rejects endogeneity of schooling at 5 percent level for TSLS 1 and accepts endogeneity in case of TSLS 2 and TSLS 3 casts doubt on the capability of the IVs and the models.

\[ e^{0.125251} - 1 = 0.133433 \]
\[ e^{0.127222} - 1 = 0.135669 \]
\[ e^{0.168077} - 1 = 0.1830277 \]

\[ e^{-0.295730} - 1 = -0.256011722 \]
\[ e^{-0.145401} - 1 = -0.135324511 \]
\[ e^{0.038858} - 1 = 0.03962295 \]
References


Appendices

Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>N=540</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
<th>RSD</th>
<th>Skew</th>
<th>Ex. K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2.78</td>
<td>2.77</td>
<td>0.82</td>
<td>4.82</td>
<td>0.63</td>
<td>0.23</td>
<td>0.10</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>13.70</td>
<td>13</td>
<td>0</td>
<td>20</td>
<td>2.48</td>
<td>0.18</td>
<td>0.27</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>wexp</td>
<td>16.67</td>
<td>17.44</td>
<td>1.77</td>
<td>23.56</td>
<td>4.72</td>
<td>0.28</td>
<td>-0.82</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>0.50</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0.50</td>
<td>1.00</td>
<td>0</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>ethblack</td>
<td>0.11</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.32</td>
<td>2.78</td>
<td>2.42</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>ethhisp</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.22</td>
<td>4.28</td>
<td>4.04</td>
<td>14.34</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>40.84</td>
<td>41</td>
<td>37</td>
<td>45</td>
<td>2.30</td>
<td>0.06</td>
<td>0.19</td>
<td>-1.08</td>
<td></td>
</tr>
<tr>
<td>married</td>
<td>0.67</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.47</td>
<td>0.70</td>
<td>-0.72</td>
<td>-1.47</td>
<td></td>
</tr>
<tr>
<td>sf</td>
<td>11.71</td>
<td>12</td>
<td>0</td>
<td>20</td>
<td>3.44</td>
<td>0.29</td>
<td>-0.37</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>sm</td>
<td>11.60</td>
<td>12</td>
<td>0</td>
<td>19</td>
<td>2.62</td>
<td>0.23</td>
<td>-0.76</td>
<td>2.55</td>
<td></td>
</tr>
<tr>
<td>siblings</td>
<td>3.25</td>
<td>3</td>
<td>0</td>
<td>13</td>
<td>2.10</td>
<td>0.65</td>
<td>1.33</td>
<td>2.60</td>
<td></td>
</tr>
</tbody>
</table>

Earnings vs Log Earnings

Q-Q Plot and Residual Plot TSL 3
Variance-Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>sf</th>
<th>sm</th>
<th>siblings</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>6.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sf</td>
<td>3.45</td>
<td>11.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sm</td>
<td>2.74</td>
<td>5.62</td>
<td>6.85</td>
<td></td>
</tr>
<tr>
<td>siblings</td>
<td>-1.07</td>
<td>-1.92</td>
<td>-1.48</td>
<td>4.40</td>
</tr>
</tbody>
</table>

TSLS and OLS Regression Results

\[
\begin{align*}
\ln Y & = \alpha_0 + \alpha_1 X_s + \alpha_2 W_{\text{wexp}} + \alpha_3 W_{\text{female}} + \alpha_4 W_{\text{ethblack}} + \varepsilon \\
\end{align*}
\]

<table>
<thead>
<tr>
<th>IVs for s</th>
<th>sf</th>
<th>sm</th>
<th>siblings</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma_{\text{OLS}}^2) &amp; 0.26 &amp; 0.26 &amp; 0.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sigma_{\text{IV}}^2) &amp; 0.28 &amp; 0.26 &amp; 0.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{Var}(\beta_{\text{OLS}})) &amp; 0.00 &amp; 0.00 &amp; 0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{Var}(\beta_{\text{IV}})) &amp; 0.00 &amp; 0.00 &amp; 0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R_P^2) &amp; 0.16 &amp; 0.17 &amp; 0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^9\) (p-values in parentheses)