# Estimating the Economic Returns to Schooling: Restricted Maximum Likelihood Approach

**Adelaide Agyeman** [1] | *Crops Research Institute, Kumasi, Ghana*
**Nicholas Nsowah-Nuamah** [2] | *Kumasi Polytechnic, Kumasi, Ghana*

**Abstract**

The economic returns to schooling is a fundamental parameter of interest in many different areas of economics and public policy. The most common technique for estimating this parameter is based on the assumption that the 'true' coefficient of education in the earnings equation is constant across individuals. However, this may not often be wholly true and returns to schooling estimates may be biased and inconsistent. The objective of this study was to estimate the returns to schooling as a random coefficient and obtain accurate and reliable estimates that will be useful for policy recommendations. The restricted maximum likelihood (REML) method was used to estimate the parameters of a random coefficient model using data from a 2007/2008 Ghanaian twins' survey. The results revealed that the REML economic returns to schooling in three selected cities were between 7% and 9%. Significant ($p<0.05$) variances around the mean returns to schooling implied that returns to schooling might vary among individuals due to unobserved factors.

| Keywords | JEL code |
|---|---|
| *Return to schooling, Random coefficient model, Maximum Likelihood, REML, Variance Heterogeneity* | *I26, J31, C1* |

## INTRODUCTION

The relationship between schooling and earnings is of key importance to the research community and policymakers in both the developed and developing countries. This is because studies have consistently confirmed that people with higher level of education earn more money, experience less unemployment, and work in more prestigious occupations than their less-educated counterparts (Card, 1999; Patrinos, 2006). An important parameter of interest frequently estimated in the schooling-earnings relationship is the economic returns to schooling. It is an indicator of schooling impact on levels of output per worker and a determinant of relative wages (Kaboski, 2007). In addition, studies of returns to schooling along

[1] Biometric Unit, Crops Research Institute, P. O. Box 3785, Kumasi, Ghana. Corresponding author: e-mail: nesiagyeman@yahoo.com, phone: (+233)322060389.

[2] Rector, Kumasi Polytechnic, P. O. Box 854, Kumasi, Ghana. E-mail: n3n_nuamah@yahoo.com.

with other research act as a guide for public policy decisions about the organization and financing of education reforms (Psacharopolous and Patrinos, 2004). In the empirical literature, the standard approach used in estimating the economic returns to schooling is the ordinary least squares (OLS) method on a simple Mincer type earnings function.

Two major issues associated with the estimation of the returns to schooling using the OLS method have been pointed out by (Card, 1999; Pfeiffer and Pohlmeier, 2012) and others. Firstly, an assumption made in most empirical studies when estimating the standard Mincerian wage equation states that the return to schooling is homogenous, (i.e., constant across individuals) making the OLS returns to schooling a fixed coefficient (i.e., a single parameter in the population). However, the return of an additional year of schooling may vary across schooling levels and across individuals of the same schooling level due to differences in observable factors (e.g. family background, school quality, level of schooling, etc.) as well as unobservable factors (e.g. cognitive and non-cognitive skills, peer group and network effects), Pfeiffer and Pohlmeier (2012). In such a situation, it may be better to regard the returns to schooling as a random coefficient subject to random variation (Hildreth and Houck, 1968). If this random coefficient is correlated with the schooling variable or the additive error term in the earnings equation, then standard OLS estimates of returns to schooling will be biased and inconsistent. Secondly, in the presence of nested and hierarchically structured data, such as individuals or twins within families, OLS techniques violate the assumption of independence of errors leading to imprecise parameter estimates and loss of statistical power, and subsequently increases the likelihood of rejecting a true null hypothesis (Raudenbush and Bryk, 2002).

Consequently, given these limitations an OLS estimation of schooling on earnings will fail to accurately identify the schooling earnings relationship and its usefulness with respect to policy recommendations will be limited. A number of economists have used the instrumental variable (IV) approach (Heckman, 1998) to address the inefficiency of OLS when returns to schooling vary across individuals. However, as noted by (Card, 2001) even the IV technique based on ideal instruments will produce estimates that are weighted averages of the returns to schooling for each individual with higher weight placed on those individuals most likely to have been affected by the instrument of choice. As a result, the IV will be a biased estimate of both the average return to schooling and the return to schooling of the group affected by the instrument if returns to schooling varies across individuals. They both concluded that in several instances the IV estimates are not precise and cannot effectively estimate policy relevant parameters.

The dominant approach to the random coefficient model estimate in recent years is based on the principle of maximum likelihood (ML) estimation (Bickel, 2007). The reason being that when the assumptions of independence of observations and residuals are violated as in the case of varying parameter estimates, maximum likelihood estimators provide parameter estimates that are relatively consistent, asymptotically normal and efficient (Card, 2001). However, the ML estimator of variance components in a linear model can be biased downwards because it does not adjust for the degrees of freedom lost by estimating the fixed regression coefficients. Patterson & Thompson (1971) introduced the restricted maximum likelihood (REML) estimator to address the limitations of the ML. REML in contrast to ML, adjusts for the degrees of freedom lost due to the estimation of the fixed effects parameters by maximizing the likelihood of linearly independent residual error contrasts to obtain unbiased estimates (Laird and Ware, 1982; Lindstrom and Bates, 1988). REML provides unbiased regression coefficients even with small samples by considering the number of parameters used in model estimation (Nunnally and Bernstein, 1994). Consistent with this trend, Ashenfelter and Krueger (1994) identified an income premium related to higher educational attainment by using data from an Ohio Twinsburg survey. Their REML returns to schooling estimate was about 16%. Mazumder (2004) also analyzed data from the 1979 National Longitudinal Survey (NLSY79) in the United States using the restricted maximum likelihood (REML) method.

His findings indicate that more than half the variation in the log of wages among men is due to differences in family and community background. Sadeq (2014) investigated differences in wage penalty between formal and informal employment using labor force survey data from three countries. His REML rate of return to required years of education for formal employees ranged from 7.8% to 8.4%. Likewise, Anger and Schnitzlein (2013) analyzed data from the German Socio-Economic Panel Study (SOEP), using a Restricted Maximum Likelihood (REML) model. They find substantial influence of family background on the skills of both brothers and sisters. Their sibling correlations of the personality traits range from 0.24 to 0.59 indicating that even for the lowest estimate, one fourth of the variance or inequality can be attributed to factors shared by siblings. Sibling correlations in cognitive skills were also higher than 0.50, indicating that more than half of the inequality in earnings could be explained by family characteristics.

The objectives of this paper are to (a) to estimate the return to schooling as a random regression coefficient, (b) to determine the influence of individual and family background characteristics on the returns to schooling and to (c) to decompose the variance around the mean return into family heterogeneity, individual heterogeneity and residual error.

## 1 MATERIALS AND METHODS

### 1.1 Data

There is no national twins database in Ghana and therefore primary data was collected by a team of five interviewers during a twins' survey in December 2007 and January 2008 in three cities in Ghana, namely Accra, Kumasi and Takoradi. Questionnaires were administered through face-to-face personal interviews to gainfully employed adult twins aged between 18 and 65. Twins were identified through various channels including twins registered at the twin's clubs, various work places, markets, shops, colleagues, friends, relatives, and households. In Kumasi 404 respondents were identified, whereas in Accra and Takoradi the total of 96 respondents were identified. Altogether, 500 respondents were identified. 50% of twins identified were randomly selected and interviewed giving a total of 250 respondents made up of 125 twin pairs. Out of the 250 respondents, 144 individuals were dizygotic (DZ) twins and 106 were monozygotic (MZ) twins. This data set provides a unique and rich source of information on the socio-economic characteristics (age, gender, marital status, earnings, education, family background characteristics such as sibling education, father's and mother's education etc.) of twins' in Ghana. Data analysis was performed using three samples (Pooled, Monozygotic and Dizygotic) in order to identify the comparative roles of genetics and family background as mediating influences in the returns to schooling.

### 1.2 Modeling Framework

The modeling technique used for estimating the return to schooling as a random coefficient was the hierarchical linear Model (HLM) by Raudenbush and Bryk, (2002). The multilevel characteristic of HLM captured the inherently hierarchical nature of the family-twins dataset (i.e. individuals/twins observations (level 1) nested within families (level 2)). The mean effect of education on earnings and the variance in returns around this mean were represented as fixed and random effects respectively. Observable differences in returns across individuals were controlled by the influence of siblings and family background characteristics (e.g. parental education) on earnings. Family-specific random returns were also estimated as deviations around the sample average return to schooling. An individual-specific random intercept was also introduced to control the unobserved heterogeneity which is usually interpreted as the return to an individual's innate ability or skill. The proportion of the total variation in earnings that lies "between" individuals in terms of an intra-class correlation (ICC or $\rho$) was also calculated to describe how strongly twins in the same family resemble each other.

As a first step in the HLM analysis of the returns to schooling, the ICC was determined using the unconditional or null model. The null model (contains no explanatory variables) expresses the individual-level earnings $Y_{ij}$ for the $i^{th}$ sibling/twin in the $j^{th}$ family ($i = 1, 2; j = 1, 2, k$) by combining two linked models: one at the individual level (level 1) and another at the family level (level 2) as:

Level-1 (sibling/twins-level) model is:

$$\mathrm{Y}_{ij} = \beta_{0j} + e_{ij}, \text{ where } e_{ij} \sim N(0, \sigma_e^2)\,. \tag{1.1}$$

Level-2 (family-level) model is:

$$\beta_{0j} = \beta_0 + \mu_{0j}, \text{ where } \mu_{0j} \sim N(0, \sigma_\mu^2)\,. \tag{1.2}$$

The level-1 and level-2 equations are combined into a single model equation and represented as:

$$\mathrm{Y}_{ij} = \beta_o + \mu_{0j} + e_{ij}, \text{ where } \mu_{0j} \sim N(0, \sigma_\mu^2), e_{ij} \sim N(0, \sigma_e^2)\,, \tag{1.3}$$

$$Var(\mu_{0j} + e_{ij}) = \sigma_{\mu 0}^2 + \sigma_e^2,\ Cov(\mu_{0j}, e_{ij}) = 0\,,$$

where $Y_{ij}$ refers to earnings for the $i^{th}$ sibling/twin in the $j^{th}$ family, $\beta_0$ is the overall mean, $\mu_{0j}$ is the random effect for the $j^{th}$ family and $e_{ij}$ is an individual-specific random error component with population variance $\sigma_e^2$. The intra-class correlation $\rho$ is then specified as:

$$\rho = \frac{\sigma_{\mu 0}^2}{\sigma_{\mu 0}^2 + \sigma_e^2}, \tag{2}$$

where $\sigma_{\mu 0}^2$ captures the variance in annual earnings that is due to differences between families while the $\sigma_e^2$ captures the variance in annual earnings within families.

Secondly, a two-level hierarchical linear model which involves the estimation of fixed effects, random returns to schooling coefficients, the variance components and individual and family variables to explain differences in returns to schooling across individuals can be written as:

Level 1:

$$\mathrm{Y}_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}, \text{ where } e_{ij} \sim N\left(0, \sigma_e^2\right), \tag{3.1}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \mathrm{Z}_j + \mu_{0j}, \tag{3.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \mathrm{Z}_j + \mu_{1j}, \tag{3.3}$$

where ($\gamma_{00}$ and $\gamma_{10}$) are the intercepts or overall means for ($\beta_{0j}$ and $\beta_{1j}$ ) from the second-level models, ($\gamma_{01}$ and $\gamma_{11}$) are the regression coefficients (slopes) from the second-level models, ($\mu_{0j}$ and $\mu_{1j}$) are the random effects or residuals for ($\beta_{0j}$ and $\beta_{1j}$), $X$ and $Z$ are matrices containing explanatory variables. X represents an explanatory variable for individual (twin) $i$ nested in level 2 (family) unit $j$, and Z represents an explanatory variable for level 2 (family) unit $j$.

Substituting Equations (3.2) and (3.3) into Equation (3.1) gives the combined model as:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j (\mu_{0j} + \mu_{1j} X_{ij} + e_{ij})\,, \tag{3.4}$$

where $e_{ij} \sim N\left(0, \sigma_e^2\right)$ and $\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right]$ = T,

where $\operatorname{var}\left(\mu_{0j}\right) = \tau_{00}$, $\operatorname{var}\left(\mu_{1j}\right) = \tau_{11}$ and $\operatorname{cov}\left(\mu_{0j}, \mu_{1j}\right) = \tau_{01}$.

$Y_{ij}$ is a function of the mean intercept ($\gamma_{00}$), the regression coefficient or mean slope ($\gamma_{10}$) for the first explanatory variable (e.g., education) at level 1, plus two random parameters (variation of the intercepts ($\mu_{0j}$) and variation of the slopes ($\mu_{1j}$) and the residual variation ($e_{ij}$). $X_{ij}$ and $Z_j$ are matrices containing individual and family level variables (e.g., education, age, marital status, parental education, etc.).

In the two-level HLM, the ($\gamma$'s) are the fixed effects parameter estimates that are assumed to be constant across individuals from Equations (3.2) and (3.3), $\beta_{0j}$ and $\beta_{1j}$ are the random effects parameter estimates that vary across individuals from Equation (3.1). ($e_{ij}$) is the variance of the first-level residuals from Equation (3.1) and ($\mu_{0j}$ and $\mu_{1j}$) are the variances of the second-level residuals.

The variance around the mean returns to schooling is decomposed into three components as:

$$Var\left(\mu_{0j} + \mu_{1j} + e_{ij}\right) = \sigma_{u0}^2 + \sigma_{\mu 1}^2 + \sigma_e^2,\ Cov\left(\mu_{0j}, e_{ij}\right) = 0,\ Cov\left(\mu_{1j}, e_{ij}\right) = 0,$$

where $\mu_{0j}$ is family heterogeneity (i.e., variance component common to all siblings in family $j$), $\mu_{1j}$ is individual or sibling heterogeneity (i.e., variance component unique to individual $i$ in family $j$) and $e_{ij}$ represents residual error due to measurement errors and other transient errors which are associated with earnings and age-related earnings differences.

### 1.3 Parameter Estimation

The two-level hierarchical model involves the estimation of three types of parameters, namely the fixed effects, random effects or random coefficients and the variance-covariance components. The restricted maximum likelihood (REML) estimator (Patterson and Thompson, 1971) was used to estimate the parameters. With the REML method only the variance components are included in the likelihood function and the regression coefficients are estimated in a second estimation step. The fixed effects are represented by ($\gamma_{00}$, $\gamma_{01}$, $\gamma_{10}$ and $\gamma_{11}$) in Equation (3.4) and were estimated by the generalized Least Squares (GLS), Laird and Ware, (1982) given variance-covariance estimates calculated by the REML method (Raudenbush, Bryk, Cheong and Congdon, 2001, p.7). The random coefficients are represented by ($\beta_{0j}$ and $\beta_{1j}$) in Equations (3.2) and (3.3) and were estimated by the empirical Bayes approach or the best linear unbiased prediction (BLUP) method. The variance-covariance components were estimated by REML method and they include (1) the covariance between level-2 error terms (i.e., cov($\mu_{0j}$, $\mu_{1j} = \tau_{01}$), (2) the variance in the level-1 error term (i.e.,var($e_{ij}$) = $\sigma_e^2$) and (3) the variance in the level-2 error terms (i.e., var($\mu_{0j}$, $\mu_{1j}$) = $\tau_{00}$ and $\tau_{11}$, respectively). The three parameters in the HLM were estimated using the Statistical Analysis System (SAS) model notation of the two-level HLM in Equation (3.4) specified as follows:

$$Y_j = A_j\gamma + X_j\mu_j + e_j, \quad j\text{=1, 2,}J, \tag{4}$$

where $A_j = X_jZ_j$, $A_j$ and $X_j$, and are known design matrices, $Z_j$ is the level 2 covariate, $\gamma$ is a vector of fixed effects, $\mu_j$ is a vector of random effects and $e_j$ is a vector of random errors. The random effects and the random errors are normally distributed with:

$$e_j \sim N\left(0, R_j\right),\ R_j = \sigma^2 I_{nj},\ \mu_j \sim N\left(0, G\right),\ G = \begin{bmatrix} \tau_{00} & \tau_{10} \\ \tau_{01} & \tau_{11} \end{bmatrix}.$$

The fixed effects ($\gamma$'s) were estimated using the GLS. The GLS estimator which provides weighted estimates of the second-level regression coefficients can be written as:

$$\hat{\gamma} = \left(A'\hat{V}^{-1}A\right)^{-1}\left(A'V^{-1}Y\right), \text{ where } V = \mathrm{var}(Y) = XGX' + R \ . \tag{5.1}$$

The variance of $\hat{\gamma}$ is given as:

$$\mathrm{var}(\hat{\gamma}) = (A'\hat{V}^{-1}A)^{-1} \ . \tag{5.2}$$

The random effects ($\mu$'s) were estimated using shrinkage estimators, namely the empirical Bayes method or the best linear unbiased prediction (BLUP) according to the equation below:

$$\hat{\mu} = \hat{G}X'\hat{V}^{-1}\left(Y - A\hat{\gamma}\right) . \tag{5.3}$$

The variance-covariance components ($\sigma_e^2$, $\tau_{00}$ $\tau_{01}$ and $\tau_{11}$) were estimated using the restricted maximum likelihood (REML) method. REML estimates of the variance-covariance components (G and R) were calculated by maximizing the REML log-likelihood function:

$$l_{REML}(G,R) = -\frac{1}{2}\log|V| - \frac{1}{2}\log\left|A'V^{-1}A\right| - \frac{N-p}{2}\log r'V^{-1}r - \frac{(N-p)}{2}\left[1 + \log\frac{2\pi}{(N-p)}\right], \tag{6}$$

where: $r = \mathrm{Y} - A\left(A'V^{-1}A\right)' A'V^{-1}\mathrm{Y}$ and $p = rank(A)$ .

The maximization was carried out using a ridge-stabilized Newton-Raphson algorithm (Lindstrom and Bates, 1988). Tests of hypotheses about the fixed and random effects and the variance-covariance components were carried out using an approximate t-statistics, Wald Z test and chi-square statistics (Polit, 1996; Agresti, 1990; Verbeke and Molenbergs, 2000). Statistical analyses were conducted using SAS Version 9.1.3 PROC MIXED with REML option.

## 2 RESULTS

Overall, female twins slightly outnumbered male twins by about 2.4% (Table 1). MZ twins earned more on average than DZ twins and fathers acquired more education than mothers (Table 2).

**Table 1** Total Number of Monozygotic (MZ) and Dizygotic (DZ) Twin Respondents in the Three Survey Areas

| Area | MZ | | DZ | | Total |
|---|---|---|---|---|---|
| | Male | Female | Male | Female | |
| Kumasi | 42 | 36 | 60 | 64 | 202 |
| Takoradi | 4 | 2 | 4 | 6 | 16 |
| Accra | 8 | 14 | 4 | 6 | 32 |
| Total | 54 | 52 | 68 | 76 | 250 |

**Source:** GTS authors' calculation

**Table 2** Descriptive Statistics – Means and Standard Errors

| Variable | Pooled sample | Monozygotic twins | Dizygotic twins |
|---|---|---|---|
| Own education (years) | 12.576<br>(0.343) | 14.009<br>(0.535) | 11.521<br>(0.427) |
| Co-twins education (years) | 12.692<br>(0.345) | 13.840<br>(0.550) | 11.847<br>(0.429) |
| Male (proportion) | 0.488<br>(0.032 | 0.509<br>(0.049) | 0.472<br>(0.042) |
| Age (years) | 32.816<br>(0.649) | 31.887<br>(0.905) | 33.500<br>(0.907) |
| Married (proportion) | 0.432<br>(0.031) | 0.321<br>(0.046) | 0.514<br>(0.042) |
| Mother's education | 5.776<br>(0.408) | 6.189<br>(0.638) | 5.472<br>(0.530) |
| Father's education | 8.288<br>(0.462) | 9.557<br>(0.723) | 7.354<br>(0.591) |
| Log of annual income | GH¢7.184<br>(0.054) | GH¢7.368<br>(0.084) | GH¢7.049<br>(0.068) |
| Sample size | 250 | 106 | 144 |

**Note:** Standard errors in parentheses below means.
**Source:** GTS authors' calculation

### 2.1 The null model or unconditional model

Table 3 represents the parameter estimates and standard errors for the null model of Equation (1.3). Results of this model reveal that the fixed effects intercept terms are approximately 7.18, 7.37 and 7.05 for pooled, MZ and DZ twins, respectively. The variance of the twins-level residual errors denoted by $\sigma_e^2$ is estimated as 0.1544. Likewise, the variance of the family-level residual effect denoted by $\sigma_\mu^2$ is estimated as 0.5714. All the parameter estimates are positive and the Wald Z-test indicates that they are also significant. The proportion of variance (i.e., the intra-class correlation coefficient (ICC)) in annual earnings that occurs between families for the pooled sample of twins is calculated as $p = 0.5714/(0.5714 + 0.1544) = 0.787$. This estimate which is very high tells us that about 80% of the total variation in earnings of twins can be accounted for by family background effect. Moreover, the ICC estimates (0.88 and 0.70) for MZ and DZ twins respectively (Table 3) indicate that about 12% and 30% of the variances in the two models are attributable to individual traits of MZ twins and DZ twins, respectively. These estimates show the extent to which observations are related within each family and therefore suggest that MZ twins are more closely genetically related than DZ twins. Overall, the correlations describe the proportion of variance associated with differences between families and indicate that family background effects contribute a sizable percentage of the variation in the returns to schooling for twins than individual effects.

Furthermore, the results of the ICC (which are greater than 10% of the total variance in the model) indicate that the HLM is an appropriate model for the estimation of the regression relationship that varies by family using multiple level data (siblings/twins nested within families, Table 3). The residual variance for all three samples are significant ($p<0.01$) and therefore supports the alternative hypothesis that average annual earnings may vary across individuals or twins with the same level of schooling.

**Table 3** Results from the Null Hierarchical Linear Model (HLM)

| **Fixed Effects** | Pooled | MZ twins | DZ twins |
|---|---|---|---|
| Family intercept, $\gamma_{00}$ | 7.1846** | 7.3686** | 7.0492** |
| | (0.0720) | (0.1158) | (0.0889) |
| **Random Effects** | Variance components | | |
| Family mean, $\tau_{00}$ | 0.5714** | 0.6639** | 0.4687** |
| | (0.0830) | (0.1396) | (0.0969) |
| Residual effect, $\sigma_{\varepsilon}^{2}$ | 0.1544** | 0.0925** | 0.2000** |
| | (0.0195) | (0.0180) | (0.0333) |
| ICC, $\rho$ | 0.7873 | 0.8777 | 0.7009 |
| **Model Fit** | | | |
| −2 Res log likelihood | 510.9 | 194.7 | 304.1 |
| AICc | 514.9 | 198.7 | 308.1 |
| N | 250 | 106 | 144 |

**Note:** * = $p < .05$, ** = $p < .01$. Standard errors in parentheses below means.
**Source:** GTS authors' calculation

The results of a second model which includes some demographic characteristics such as number of years spent schooling, age, gender, marital status, father's education and mother's education as explanatory variables are presented in Table 4. Average annual earnings (5.68, 5.02 and 5.88) for the pooled, MZ and DZ twins' samples respectively, are highly significant ($p<0.01$), suggesting that the effects of education on earnings vary from one family to another and among individuals. The findings also indicate that data used is dominated by a hierarchical structure, which may affect both the intercepts and the slopes (returns to education) of earnings functions. The results further indicate that expected earnings for the three data sets were similar irrespective of the type of model (null or random coefficient) used. However, the expected earnings of the different groups were higher for the null model compared to the second model. Apparently, accounting for the variation in sibling earnings by including demographic variables decreases expected earnings (intercepts) by about 1.5 and 1.2 points for MZ and DZ twins respectively when compared to the expected earnings of the null model which did not have any covariates. This suggests that demographic characteristics explain a proportion of the variation in annual earnings. The effect of an additional year spent schooling on individual earnings ranged from 7% to 9% for the three data samples (Table 4) and it differed significantly from zero (i.e., $p<0.01$). Returns to schooling estimates for MZ twins were lower than that of both the pooled and DZ twins. This may indicate the existence of some upward bias for MZ twins REML estimates due to omitted unobserved characteristics and also confirms the fact that failure to take account of unobserved heterogeneity leads to biased estimates on the returns to schooling. It may also suggest that high-ability MZ twins find it easier to acquire more education. Father's education significantly ($p<0.05$) affected REML returns to schooling for MZ twins, whiles mother's education had a significant impact on REML returns to schooling for DZ twins. The returns to schooling estimates (Pooled = −0.15, MZ = −0.10 and DZ = −0.16) for gender measured by the dummy male were negative for all three samples and significant at the 5% level for the pooled and DZ twins' samples. This inverse relationship implies negative average returns to education for male twins and suggests that an additional year of schooling has a higher pay-off for females than for males. This means that while females have lower wage levels than men, they have higher average returns to education. The effect of age on earnings for every additional life year was significant ($p<0.05$) for MZ twins but insignificant ($p>0.05$) for DZ twins. This finding may be associated with age being a better proxy for actual work experience for MZ twins than it is for DZ twins. Moreover, the MZ twins sample are on average younger than the DZ twins and therefore a decline in earnings could come about at older ages. The effect of the proportion of those who are married on earnings was negative and not significant (*p*>0.05) for all three data samples, indicating that being married does not guarantee an individual an increase in earnings.

### 2.2 Variation around the mean returns to schooling

Comparison of the variance components corresponding to the random intercepts (family-level variance) between the null and second models (Tables 3 and 4) shows that family-level variance components for MZ twins decreased by 70% in the second model (Table 4). This indicates that individual and family characteristics explain a larger portion of the differences in the returns to schooling for MZ twins and that an earnings-education model that does not take into account these characteristics may overestimate the returns to schooling. Although, the family level variance for MZ twins in the second model is not significantly different from zero (*p*>0.05), the variance around the mean returns to schooling is, however, significant (*p*<0.05), Table 4.

**Table 4** Results of the Hierarchical linear Model (HLM) including Covariates

| **Fixed Effects** | Pooled | MZ twins | DZ twins |
|---|---|---|---|
| Family intercept, $\gamma_{00}$ | 5.6777** | 5.0160** | 5.8847** |
| | (0.2372) | (0.2999) | (0.3070) |
| Schooling (years) slope, $\gamma_{10}$ | 0.0878** | 0.06801** | 0.0886** |
| | (0.0113) | (0.0173) | (0.0133) |
| Age (years) | 0.0141* | 0.0399** | 0.0085 |
| | (0.0061) | (0.0084) | (0.0076) |
| Gender | –0.1488** | –0.1022 | –0.1643** |
| | (0.0570) | (0.1315) | (0.0587) |
| Married | –0.0038 | –0.0286 | –0.0024 |
| | (0.0896) | (0.1368) | (0.1034) |
| Father's schooling (years) | 0.0061 | 0.0323** | –0.02051 |
| | (0.0112) | (0.0111) | (0.0169) |
| Mother's schooling (years) | 0.0137 | –0.0093 | 0.0430* |
| | (0.0130) | (0.0132) | (0.0191) |
| **Random Effects** | Variance Components | | |
| Family variance, $\tau_{00}$ | 0.7652** | 0.1835 | 0.9521** |
| | (0.2105) | (0.2632) | (0.2884) |
| Schooling slope, $\tau_{11}$ | 0.0031** | 0.0031* | 0.0028** |
| | (0.0012) | (0.0019) | (0.0011) |
| Residual effect, $\sigma_{\varepsilon}^{2}$ | 0.07456** | 0.0827** | 0.0698** |
| | (0.0102) | (0.0163) | (0.0135) |
| **Model Fit** | | | |
| –2 Res log likelihood | 398.1 | 164.8 | 242.0 |
| AICc | 406.1 | 172.8 | 250.0 |
| N | 250 | 106 | 144 |

**Note:** * = $p < .05$, *** = $p < .001$. Standard errors in parentheses below means.
**Source:** GTS authors' calculation

The variance components for the returns to schooling coefficient in the second model is denoted by $\tau_{11}$ and estimated as 0.0031 and 0.0028 with standard errors of 0.0019 and 0.0011 for the MZ and DZ twins' respectively (Table 4). These variances are higher than their standard errors suggesting that the second model picks up most of the variance in the returns to schooling that exist across families, though this variation is still significant ($p$<0.05). The significant variances of the regression slopes for the MZ and DZ twins data imply that returns to schooling varies across families and the values of 0.07 and 0.09 are just the expected returns across families (Table 4). This indicates that there could be some level of unobserved differences between MZ twins which may be attributed to individual characteristics. Similar random coefficient variance estimates are also associated with the returns to schooling for the pooled and DZ twins datasets and are significantly different from zero ($p$<0.05) using the Wald Z-test. This shows that the returns to schooling for these twins differ more than one could reasonably attribute to chance. REML returns to schooling results from Table 4 show significant ($p$<0.05) variation in the estimated intercepts and slope coefficients and therefore suggest that there exists heterogeneity in the returns to schooling. Since the random effects for the MZ and DZ twins are assumed to follow a normal distribution, about 67% of the returns to schooling regression coefficients for the MZ twins are expected to lie between an interval of (0.0123 and 0.1237) and about 95% are predicted to lie between (0.0411 and 0.1771). Similarly, about 67% of the returns to schooling regression coefficients for DZ twins are expected to lie between (0.0357 and 0.1415) and about 95% are predicted to lie between (–0.0151 and 0.1923). Thus, a return to schooling corresponding to the lower interval would indicate that if an employee is a DZ twin, annual family earnings is decreased by approximately 1.5% when compared with returns to schooling for non-DZ twins. Likewise the returns to schooling that corresponds to the upper limit of the interval would mean that annual family earnings for a DZ twin employee increased by 19% when compared with returns to schooling for non-DZ twins. A returns to schooling corresponding to the lower interval would indicate that if an employee is a MZ twin, annual family earnings are decreased by less than 5% when compared with returns to schooling for non-MZ twins. Likewise the returns to schooling that correspond to the upper limit of the interval would mean that annual family earnings for a MZ twin employee increased by 18% when compared with returns to schooling for non-MZ twins.

Furthermore, the Wald-Z test pointed out that the residual components which measured the variation not accounted for in the hierarchical linear models for both MZ and DZ twins in the null and second models were statistically significant ($p$<0.05). Interestingly, the residual variance associated with the returns to years of schooling for MZ twins in the second model decreased by about 11% whiles that of DZ twins decreased by about 65% (Table 4) when compared to the null model residual variances. This suggests that there is still some unobserved variation in returns to schooling for both MZ and DZ twins which could be attributable to measurement error in reported schooling levels and possible individual differences in inherent ability, among other reasons. Additionally, the significant REML residual variation is essentially due to the randomness of observed rates of returns to schooling and is an indication that returns to additional schooling varies randomly across individuals due to factors unknown to both the researcher and the individual at the time of their decisions.

According to the smaller-is-better rule for the information criteria, Model 2 has a smaller (AICc) (406.1) and a lower Restricted log likelihood (–2RLL) (398.1) compared to (AICc – 514.9) and (–2RLL – 510.9) of the null model and is therefore considered the best model. The probability chi-square of the difference in the log likelihood test of the models for the MZ, DZ and Pooled data sets, revealed that there were significant ($p$<0.01) differences between the null and the second model with explanatory variables (Table 5).

**Table 5** Testing the significance of 2 Hierarchical linear Models for MZ and DZ Twins

| Item | Difference in Log likelihood (–2LL) | Difference in *df* | *p*>chi-square |
|---|---|---|---|
| MZ – Model 1&2 | 29.9 | 2 | 3.21586E-07 |
| DZ – Model 1&2 | 62.1 | 2 | 3.27459E-14 |
| Pooled – Model 1&2 | 112.8 | 2 | 3.20473E-25 |

**Source:** GTS authors' calculation

## 3 DISCUSSION

Returns to schooling have been estimated as fixed coefficients using OLS methods in a number of labor economics studies. However, the OLS estimates may be inconsistent and biased when the returns to schooling vary across individuals as a result of observable factors as well as unobservable factors. Card, (1995) observed in a number of studies that different individuals acquire different returns to schooling and the same individual's returns to schooling vary with the level and type of schooling. In such situations the assumptions of non-varying slopes and intercepts, and uncorrelated residuals in standard OLS estimates are violated. Maximum likelihood estimators address the violation of the assumption of fixed coefficients by permitting intercepts and slopes to vary from one group to another. Moreover, in real life situations data collected are mostly of a hierarchical nature and statistical measures must be taken to exploit the opportunities offered by multilevel data structures. In order to obtain efficient and consistent estimates of the returns to schooling for a set of fully employed MZ and DZ twins, we used the REML estimation procedure in a hierarchical linear model to estimate the returns to schooling. Three types of parameters, namely the fixed effects, random effects and the variance-covariance components were estimated. REML estimated an unbiased variance around the mean returns to schooling parameters by accounting for the degrees of freedom lost by the estimation of the mean returns to schooling.

The estimated rates of return to schooling for the pooled, MZ and DZ twins ranged between 7% and 9%. These rates of returns to schooling are comparable to that of Conneely and Uusitalo (1999) who estimated a random coefficient model using Finish data that allowed for endogenous schooling and ability bias with an estimated maximum likelihood mean return to schooling of 6%. Similarly, the REML returns to schooling estimates of Sadeq (2014) using a hierarchical linear model varied between 7.8% and 8.4%. Moreover, Ashenfelter and Krueger (1994) found a higher restricted maximum likelihood estimate (16%) of the returns to schooling. Altogether, these results provide consistent and efficient estimates of the returns to schooling across individuals. The positive and somewhat large returns to schooling in the hierarchical linear model also indicate the importance of accounting for unmeasured ability and motivational factors that affect the returns to schooling.

Interestingly, MZ twins' earnings were significantly affected by fathers' education whiles mothers' education significantly influenced DZ twins' earnings. Thus, MZ twins' had better educated fathers who increased their children's education through transmission of innate ability, whereas DZ twins' had better educated mothers who raised their children's education by enhancing the "family learning environment." The effect of family background characteristics, i.e., parental education on returns to education is an important topic in the economics literature (Griliches, 1979). Part of this importance stems from the strong correlation between the educational attainment of parents and children, which may contribute to the transmission of socioeconomic status and inequality across generations. Parental education was found to positively and significantly affect the earnings of both MZ and DZ twins. Similarly, Anger and Schnitzlein (2013) concluded that family background variables play an important role in generating variation in the return to schooling. However, using twins data, Ashenfelter and Rouse (1998) are of the view that the effects of family background (and ability) on returns are small. Altonji and Dunn (1996) also measured the effects of family background on the returns to schooling and found a positive though small effect of family background on returns in their preferred fixed effects specification.

Observed rates of returns to education may vary across individuals within the same educational group because of risk and unobserved heterogeneity. This study therefore added individual and family factors to the HLM to account for some of the variation in returns to schooling. The variance around the mean returns to schooling was decomposed into family heterogeneity, individual heterogeneity and risk. Significant ($p$<0.05) individual differences in the variance around the mean returns to schooling were observed for both MZ and DZ twins. However, in contrast to DZ twins, there were no significant unobservable family differences around the mean returns for MZ twins. This confirms the fact that MZ twins have similar ability and similar family background. This is consistent with findings by Ashenfelter and Krueger (1994) and Yew (2000) who did not find any statistically significant (i.e., $p$<0.05) sources of heterogeneity in the returns to schooling for MZ twins. These MZ twins' results suggest that individuals from higher ability families receive a lower marginal benefit from their human capital investment. On the contrary, significant family heterogeneity for DZ twins indicate that able individuals may attain more schooling because of higher marginal benefits to each additional year of education. Similarly, Bingley et al. (2005) exploited panel data using mixed model to show that there were significant variances to the returns to schooling estimates and found that individual variance in returns is smaller for MZ twins than for DZ twins. Correspondingly, Chen (2002) used US panel data (NLSY) to separate the variation in the returns to college into heterogeneity and risk components, and found that almost all the variation in returns is accounted for by the heterogeneity component.

Investing in education is always associated with some amount of risk (Hartog, 2011). This risk is the variation in the returns to education due to factors unobserved by the individual. The residual variance estimate which represents individual earnings risk was about 8% for MZ twins and 7% for DZ twins. These estimates are in line with some of the existing literature. Koop and Tobias (2004) apply the model to the NLSY and find a mean return of 12% with a dispersion of 7%. Chen (2002) also finds that the dispersion in returns to a US college education is 7%. Thus, the risk is quite large, even though we have allowed for differences by observable characteristics and it implies that a large number of the twins data set show very low returns to education. Interestingly, the residual variance for both MZ and DZ twins were statistically significant suggesting that the earnings risk associated with an additional year of schooling is important and therefore needs policy interventions. This earnings risk may result from lack of knowledge about individual ability and unanticipated changes in market conditions.

## CONCLUSION

In this paper we have examined the restricted maximum likelihood (REML) estimation of a random coefficient model for earnings and its potential to provide unbiased returns to schooling regression coefficients. Results from our statistical and econometric analysis show that the mean return to schooling in the three selected cities in Ghana is between 7% and 9% which is comparable with worldwide estimates. Using the REML approach, the study also observed that there were significant variations around the mean returns to schooling across individuals which may partly be due to unobservable differences in individual ability and family background characteristics. The study further observed that family background characteristics (i.e. parental education) positively and significantly affect the earnings of both MZ and DZ twins. This is an indication that family background characteristics may play an important role in the relationship between earnings and schooling for genetically identical and similar twins. Consequently, the REML approach provides a robust alternative to the ordinary least squares method when returns to schooling vary across individuals and when data used is hierarchically structured. REML approach to the estimation of the returns to schooling offers a measure of the true effect of schooling on earnings which has important implications for policy formulation and decision making within the education sector especially for developing countries.

## ACKNOWLEDGEMENTS

We are grateful to the late Professor Louis Muyankazi of the Department of Mathematics and Statistics of the Kumasi Polytechnic and the academic staff of the Department of Mathematics and Statistics of the University of Cape-Coast, Ghana for their comments.

## *References*

AGRESTI, A. *Categorical Data Analysis:* New York: John Wiley & Sons, Inc., 1990.

ALTONJI, J. G. AND DUNN, T. A. The Effects of Family Characteristics on the Return to Education. *Review of Economics and Statistics,* 1996, 78, pp. 692–704.

ANGER, S. AND SCHNITZLEIN, D. D. *Like brother, like sister? The importance of family background for cognitive and non-cognitive skills.* Annual Conference (Duesseldorf): Competition Policy and Regulation in a Global Economic Order, Verein für Socialpolitik/ German Economic Association, 2013.

ASHENFELTER, O. AND KRUEGER, A. Estimates of the economic return to schooling from a new sample of twins. *American Economic Review*, 1994, 84, pp. 1157–1173.

ASHENFELTER, O. AND ROUSE, C. Income, Schooling and Ability: Evidence from a New Sample of Identical Twins. *The Quarterly Journal of Economics*, 1998, 113, pp. 253–284.

BICKEL, R. *Multilevel analysis for applied research: It's just regression!* New York: Guilford Press, 2007.

BINGLEY, P., CHRISTENSEN, K., WALKER, I. *Twin-based estimates of the returns to education: evidence from the population of Danish twins*. Mimeo, 2005.

CARD, D. Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, 2001, 69, pp. 1127–1160.

CARD, D. The causal effect of education on earnings. *Handbook of Labor economics*, 1999, 3, pp. 1801–1863.

CARD, D. Earnings, schooling, and ability revisited. *Research in Labor Economics*, 1995, 14, pp. 23–48.

CHEN, S. *Is investing in college education risky?* University of Rochester and University of New York at Albany, Mimeo, 2002.

CONNEELY, K. AND UUSITALO, R. *Estimating Heterogeneous Treatment Effects in the Becker Schooling Model*. Princeton University, Mimeo, 1998.

GRILICHES, Z. Sibling models and data in economics: Beginning of a survey. *Journal of Political Economy*, 1979, 87, pp. 37–64.

HARTOG, J. Allocation and the earnings function. *Empirical Economics*, 1986, 11, pp. 97–110.

HECKMAN, J. J. AND VYTLACIL, E. J. Instrumental Variables Methods For the Correlated Random Coefficient Model: Estimating The Average Rate of Return to Schooling When the Return Is Correlated With Schooling. *Journal of Human Resources*, 1998, 33, pp. 974–1002.

HILDRETH, C. AND HOUCK, J. Some Estimators for a Linear Model with Random Coefficients. *Journal of the American Statistical Association*, 1968, 63, pp. 584–595.

KABOSKI, J. *Explaining schooling returns and output levels across countries*. Unpublished manuscript, 2007.

KOOP, G. AND TOBIAS, J. *Learning about Unobserved Heterogeneity in Returns to Schooling.* Dept. of Economics, University of Glasgow, Mimeo, 2002.

LAIRD, N. M. AND WARE, J. H. Random-effects models for longitudinal data. *Biometrics*, 1982, 38, pp. 963–974.

LINDSTROM, M. J. AND BATES, D. M. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association*, 1988, 83, pp. 1014–1022.

MAZUMDER, B. Sibling similarities and economic inequality in the US. *Journal of Population Economics*, 21, 2008, pp. 685–701.

NUNNALLY, J. AND BERNSTEIN, I. *Psychometric Theory*. New York: McGraw-Hill, 1994.

PATTERSON, H. D. AND THOMPSON, R. Recovery of interblock information when block sizes are unequal. *Biometrika*, 1971, 58, pp. 545–554.

PATRINOS, H. A., RIDAO-CANO, C. AND SAKELLARIOU, C. *Heterogeneity in Ability and Returns to Education: Multi-country Evidence from Latin America and East Asia*. World Bank Policy Research Working Papers, No. 4040, Washington, D.C., 2006.

PFEIFFER, F. AND POHLMEIER, W. *Causal returns to schooling and individual heterogeneity.* IZA Discussion Paper, No. 6588, 2012.

POLIT D. *Data Analysis and Statistics for Nursing Research.* Stamford, Connecticut: Appleton & Lange, 1996.

PSACHAROPOULOS, G. AND PATRINOS, H. Returns to investment in education: A further update. *Education Economics*, 2004, 12, pp. 111–134.

RAUDENBUSH, S. W. AND BRYK, A. S. *Hierarchical linear models: Applications and data analysis methods*, 2$^{nd}$ Ed. Newbury Park, CA: Sage Publications, 2002.

RAUDENBUSH, S., BRYK, A., CHEONG, Y. F. AND CONGDON, R. *HLM 5: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, 2001.

SADEQ, TAREQ. *Formal-Informal Gap in Return to Schooling and Penalty to Education-Occupation Mismatch a Comparative Study for Egypt, Jordan, and Palestine*. Working paper series, 894, 2014.

VERBEKE, G. AND MOLENBERGS, G. *Linear Mixed Models for Longitudinal Data*. NY: Springer, 2000.

YEW LIANG LEE. Optimal Schooling Investments and Earnings: An Analysis Using Australian Twins Data. *The Economic Word*, 2000, 76, pp. 225–235.