Economics Bulletin

Volume 36, Issue 3

Schooling, experience and earnings: international evidence from MARS

Geraint Johnes
Lancaster University

Abstract

The method of Multivariate Adaptive Regression Splines is applied to international data to evaluate earnings functions. This reveals interaction effects not highlighted in more traditional analyses. In particular, a small number of nonlinear country-specific effects in the relationship between education and earnings is identified.

Without implication, the author acknowledges the helpful comments of two referees and an editor of this journal.
Citation: Geraint Johnes, (2016) "Schooling, experience and earnings: international evidence from MARS", *Economics Bulletin*, Volume 36, Issue 3, pages 1287-1294
Contact: Geraint Johnes - G.Johnes@lancs.ac.uk.
Submitted: March 10, 2016. **Published:** July 08, 2016.

1. Introduction

The functional form of wage equations has been well established in the literature since the seminal work of Mincer (1974). The logged wage is modelled as a function of linear and quadratic terms in experience (sometimes proxied by potential experience or age), of education (measured by years or by attainment), and of a variety of other explanatory variables. While this nonlinearity in the relationship between experience and earnings is motivated by theory (see, for example, Ben-Porath, 1967), the specific type of nonlinearity is not; indeed the quadratic form has been challenged in the literature (see, for example, Murphy and Welch, 1990). That said, the polynomial terms in experience are appealing not least because (at least when higher order terms are included) they fit the data well, consistently revealing a relationship where experience augments wages at a decreasing rate, often with wages peaking shortly before retirement. The polynomial specification is further appealing because it represents a Taylor approximation to more general processes, including neural networks which may themselves be regarded as universal approximators (Hornik *et al*., 1989).

In this note, I pursue an alternative route to universal approximation, and apply this to microdata drawn from more than 30 countries. This allows both commonalities and differences across countries to be explored in a consistent framework; the comparative nature of the exercise may furthermore suggest explanations that might underpin any differences that emerge. The approach taken here turns out to be particularly instructive in highlighting idiosyncrasies in the way that some large economies reward educational attainment.

The paper proceeds as follows. Section 2 concerns the methodology used. The data are introduced in Section 3. This is followed by the empirical analysis in Section 4. Conclusions are drawn in Section 5.

2. Methodology

The method of Multivariate Adaptive Regression Splines (MARS) was introduced by Friedman (1991) as a means of conducting a flexible non-parametric regression analysis designed to fit the data accurately. It is a method particularly useful in contexts where, as here, the dimensionality of the problem is moderate to high, in that it provides a means of identifying important interactions between explanatory variables without introducing so many that overfitting becomes a problem. These interactions include those between any given explanatory variable and itself – in other words, features characteristic of a polynomial are accommodated in the model, in this instance by flexibly introducing splines.

While simple in principle, the practical application of this type of non-parametric approach is hindered by the curse of dimensionality (Bellman, 1961). The MARS method provides an algorithm that allows splines and interactions to be introduced in the model while providing criteria that prevent their proliferation.

MARS models are built up from (up to) three types of basis function. These are: (i) a constant; (ii) a hinge function, taking the form $\max(0, x - \beta)$ or $\max(0, \beta - x)$ where x is an explanatory variable and β a knot (that is, a constant at the kink of a spline, this constant being estimated by the statistical procedure), and (iii) the product of up to d hinge functions, where d is the maximum degree of interaction specified by the analyst (in the present case, 2).

The presence of product terms of this kind allows smooth nonlinearities in the relationship between the dependent and explanatory variables in a MARS (Friedman, 1999, pp.13-14). The basis functions comprising a MARS are aggregated, using weights determined statistically, to form the full model.

The problem is solved by following an algorithm until a termination criterion is met. This algorithm comprises a 'forward' pass and a 'backward' pass. The constant (intercept) term is included in all stages.

In the forward pass, the model is augmented by introducing, sequentially, pairs of hinge functions (and hinge interactions) – choosing amongst all explanatory variables and all possible knot locations. At each step, the pair that most reduces the residual sum of squares (RSS) is chosen and added to the 'parent' group of terms. The process of adding to the parent group is continued until the termination condition is satisfied. This condition is defined by the analyst; in the present case, it is set so that, during the forward pass, R^2 should increase by less than 10^{-5}.

The backward pass of the algorithm is then applied in order to prevent overfitting. In this pass, terms are deleted one by one to minimise the generalised cross validation (GCV) statistic (Friedman, 1991, p.20), where

$$GCV = RSS / N(1-m(1+p/2)/N+p/2N)^2$$

Here RSS denotes the residual sum of squares, N is the number of observations, m is the number of knots, and p is a penalty term, which, following Friedman (1991, p.21), I set equal to 3. Note that the penalty interacts with the number of knots in the denominator of this expression so that greater model complexity is penalised. Consequently the backward pass of the MARS algorithm favours parsimony in the model.

MARS models have the advantage of flexibility over standard regression methods, but like other non-parametric techniques suffer the drawback that confidence intervals cannot be routinely calculated. Since we know from thousands of studies conducted after Mincer (1974) that the form of the wage equation is nonlinear, the benefit of flexibility seems to be especially worth pursuing in this instance.

The estimates reported later in this note are obtained using the earth package in R.

3. Data

The data used in this study come from the 2012 International Social Survey Programme (ISSP). These are freely available online at http://www.issp.org/. The ISSP is an internationally coordinated suite of national surveys of individuals. The surveys have been conducted annually since 1985. The first of these gathered data from Austria, Australia, Germany, Italy, the United Kingdom and the United States. The geographical scope of the surveys has increased significantly over time, and in 2012 covered some 40 countries. The range of countries represented is wide: The United States and Canada are included, as are most European Union countries and most of the BRICS. Countries are selected into the ISSP on the basis of their willingness to participate; there is no criterion in terms of a minimum level of economic development, and the countries included in the analysis differ considerably

from one another in this respect. This is a major advantage of the ISSP; it allows comparison across countries that are very different from one another.

In each year, the ISSP focuses on a particular theme of interest to social scientists. In 2012 this theme concerned the Family and Changing Gender Roles. The surveys in each year include, however, also a core of demographic about respondents that is useful for the study of their labour market experiences.

Owing to the need to collate the information from numerous national statistical agencies, microdata are released to researchers with a lag. The latest data currently available are for 2013. I choose to use instead the 2012 data because the sample used in this year is much larger than in the later year – across all countries in the study, there are some 56254 observations available in 2012, but only 25031 in 2013. Owing to our interest in the labour market, and specifically in those individuals who are earning, not all respondents are included in our analysis. Any analysis at country level would therefore be compromised by small sample size in the 2013 data.

4. Analysis

The dependent variable used in the analysis that follows is the log hourly income of the respondent, calculated from income data and information about hours worked. This is measured in local currency, and a full set of country dummy variables is included in the analyses that follow in order to allow for this. The explanatory variables include years of education, age[1], and dummy variables indicating gender (male = 1) and whether the respondent's job includes supervisory responsibilities.[2] Estimates are for data pooled across countries and so a full set of country dummies is also included.[3] The sample of respondents comprises the 12282 employees with positive income working at least 10 hours per week and aged between 20 and 60.[4]

Descriptive statistics are reported in Table 1. The average age of the sample is close to 40 years, and the average years of completed education number about 14. Just over a quarter of respondents work in a job which involves some supervisory responsibility.

---

[1] Age is used as a proxy for experience, since the ISSP does not provide information on years worked. This may have implications for the analysis, particularly if there is cross-country variation in the extent to which interrupted careers cause, for some workers, a gap to arise between true experience and a measure of potential experience (given by age minus education). In the MARS analysis, however, the possible interaction of age, education, gender and country terms means that this is allowed for in the analysis.

[2] A referee has questioned the inclusion of this last variable on the grounds that it may be capturing some of the effect of education on earnings. In practice, however, the correlation between supervision and education is very low, at 0.04.

[3] The countries in the study are: Argentina; Austria; Australia; Bulgaria; Canada; Chile; China; Taiwan; Croatia; Czech Republic; Finland; France; Germany; Iceland; India; Ireland; Israel; Japan; Korea; Latvia; Lithuania; Mexico; Norway; Philippines; Poland; Russia; Slovakia; Slovenia; Spain; Sweden; Switzerland; Venezuela; the United Kingdom; the United States. Since an intercept term must be included in the MARS, one country dummy (Argentina) must be omitted from the analysis to avoid multicollinearity. A small number of countries for which data are available in the ISSP are omitted because data on the variables of interest in the present study are incomplete.

[4] The sample used in the analysis that follows typically includes several hundred observations within each country. The largest country sample is 693 (Taiwan). The samples for India and the Philippines are marked exceptions to the general rule and are very small (81 and 78 respectively).

Table 1 Descriptive statistics

Variable	Mean	Standard deviation
Logged earnings	3.87	2.02
AGE	39.93	10.08
EDUCYRS	14.33	2.46
SEX	0.47	0.50
SUPERVISOR	0.26	0.44

Note: EDUCYRS denotes years of education.

To provide a point of comparison for the MARS results that appear later in the paper, Table 2 reports the parameter estimates obtained when estimating a conventional Mincerian model with a full set of country dummies. The parameters are all highly significant and are in line with estimates observed elsewhere in the literature. The Mincerian return to education is around 7%. The wage peaks at around age 47. Since the data being used here are cross-section data, cohort effects and pure age effects may be conflated. There are premia associated with being male and with having supervisory responsibilities, in each case amounting to around 17%. The model includes country fixed effects.

Owing in part to the presence of these fixed effects, the fit of the equation is good, with an R^2 value of 0.921. It is important to note that this sets a high hurdle in the MARS analysis that follows, since, as noted above, any hinge functions included in the model must serve to increase R^2 by more than the prescribed threshold. The MARS model is therefore likely to include country-specific effects of the explanatory variables only when these contribute significantly to overall explanatory power – not, as is the case with the t statistic used in conventional regression analysis, when the coefficient estimate is precisely estimated.

Table 2 OLS estimates of the (logged) earnings function

Variable	Coefficient
Intercept	0.495
	(6.39)
AGE	0.0562
	(15.83)
AGE^2	-0.0006
	(13.13)
EDUCYRS	0.0687
	(31.80)
SEX	0.1762
	(7.48)
SUPERVISOR	0.1699
	(14.36)
R^2	0.928
Adjusted R^2	0.928

Note: t statistics in parentheses. In addition to the variables listed in the table, a full set of country fixed effects is included.

Results of the MARS analysis are reported in Table 3. Recall that confidence intervals are not available for these coefficient estimates, but that the retention in the model of the variables and hinges shown in the table confirms that they have a significant effect on earnings. The coefficient on gender implies a wage differential of just under 20 per cent. No interaction terms between gender and country dummies are selected into the model, indicating that this differential is stable across countries. The premium associated with supervisory responsibilities is around 17 per cent, and is likewise stable across countries.

Table 3 MARS estimates of the (logged) earnings function

Intercept		3.1198		h(EDUCYRS-16)		0.0271	
SEX		0.1813		h(AGE-29)*CH		0.0139	
SUPERVISOR		0.1542		h(EDUCYRS-13*CL		0.1637	
h(29-AGE)		-0.0825		h(EDUCYRS-14)*CN		0.2471	
h(AGE-29)		0.0037		h(16-EDUCYRS)*US		-0.1083	
h(16-EDUCYRS)		-0.0710		h(30-AGE)*h(16-EDUCYRS)		0.0113	
GCV	0.2937	RSS	3551	Generalised R^2	0.928	R^2	0.929

Note: In addition to the variables listed in the table, a full set of country fixed effects is included. h refers to a hinge. The country codes CH, CL, CN and US refer respectively to Switzerland, Chile, China and the United States. Conventional standard errors cannot be computed for this type of model; variables are included in the model only if they satisfy the requirement imposed by the termination condition.

Age and education both impact on log wages nonlinearly – and the nature of that nonlinearity is complex. There is a knot in age at 29, beyond which point the relationship flattens considerably. There is a knot in years of education at 16, beyond which the relationship likewise flattens somewhat. There is also an interaction between education and age that indicates a dampening of the wage premium for older, more highly educated workers. This term introduces a continuous nonlinearity in the relationships between the wage and both age and education.[5]

A small number of interaction terms between country dummies and the age and education variables are retained in the model. Two of these are particularly worthy of note. They refer to the relationship between education and wages in China and in the United States. In both countries the slope of the relationship between education and earnings differs from that observed elsewhere, the gap between those at the top of the education distribution and those at the bottom being bigger. The reason for this differs markedly across the two countries, however. In China, those respondents with high levels of education are rewarded considerably more than in other countries for marginal increases in their education, while the return to education at lower levels is in line with that observed elsewhere. In the United States, meanwhile, returns to education at high levels of education are commensurate with

[5] As noted earlier, the criterion for the inclusion of interaction terms in the MARS is based on overall fit, not on the value or precision of the coefficients on the interaction terms themselves. It is, however, of interest to note the range of estimates obtained on key explanatory variables in separate country-specific OLS regressions. For years of education, the lowest coefficient observed is 0.04 (in several countries); the highest (in the Philippines) is 0.34, and the second highest (in Chile) is 0.18. Most coefficients are in the range 0.05-0.10. For gender, the coefficient ranges from a small and statistically insignificant negative value in India to 0.51 in Japan. For supervisory activity, the coefficient ranges from a small and statistically insignificant negative value in India to 0.33 in Korea. The country-specific estimates reported here should be treated with caution in view of the limited sample sizes in the case of some countries.

those observed elsewhere, but the rate of return for those achieving only low levels of education are lower. One might speculate that the high premium for highly educated workers in China likely reflects the scarcity of highly qualified workers in the context of a rapidly developing economy, while the low premium at the bottom end of the labour market in the United States may result from global competition (Freeman, 1995).

The main features of the results reported above can usefully be illustrated graphically. Figures 1 shows the nonlinear relationships between age, education and earnings, while Figures 2 and 3 respectively show the idiosyncratic impact of education on earnings in China and the United States. In Figure 2, it is seen that being in China (a binary move up the CN axis) results in a steepening of the relationship between years of education and earnings as education increases. In Figure 3, it is seen that being in the US (a binary move up the US axis) results in a fall in the earnings return to education at low levels of education, though this rapidly rises toward the norm observed in other countries as years of education increase.

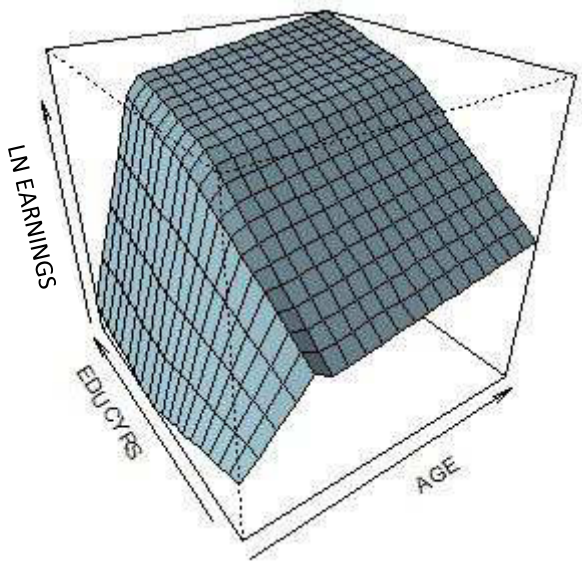

Fig. 1. Earnings, age, and years of education

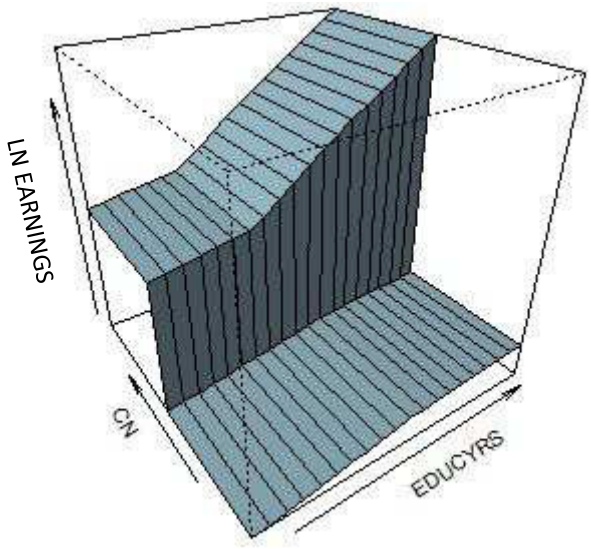

Fig. 2. Earnings and years of education in China

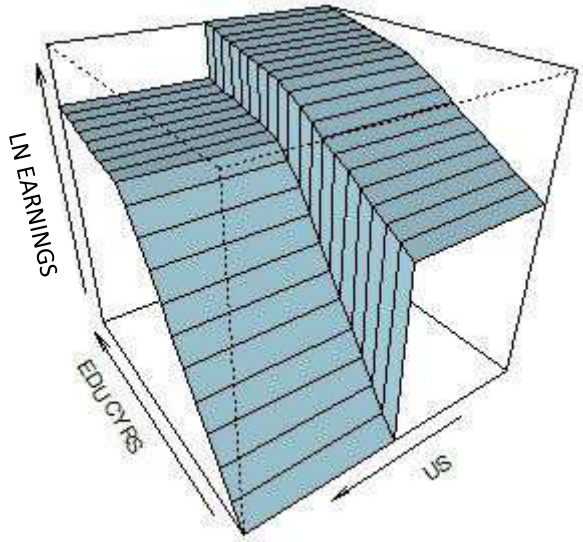

Fig.3. Earnings and years of education in the United States

Both the standard Mincerian model and the MARS model have high explanatory power, and many findings are common to both models. The gain in complementing the standard model with a MARS analysis is evident in the identification of country-specific peculiarities. In this instance, it has highlighted the unusually low return on education in the US at low levels of education and the unusually high return in China at high levels of schooling. In these respects, these two countries differ from what is typically observed in the many other countries in our data.

5. Conclusion

The application of MARS methods to the case of earnings functions confirms much that we already knew about, for instance, the nonlinear relationship between age and earnings. In the present exercise, the analysis has been conducted using international data, and this has

highlighted a small number of peculiarly interesting differences across countries. The different sources of high returns to education in China and the United States – whereby, in the former country, returns to higher education are particularly high, whereas, in the latter country, returns to early education are unusually low – seem, in particular, to be worthy of further research.

References

Bellman, Richard E. (1961) *Adaptive Control Processes*, Princeton: Princeton University Press.

Ben-Porath, Yoram (1967) The production of human capital and the life cycle of earnings, *Journal of Political Economy*, 75, 352-365.

Freeman, Richard B. (1995) Are your wages set in Beijing?, *Journal of Economic Perspectives*, 9(3), 15-32.

Friedman, Jerome H. (1991) Multivariate Adaptive Regression Splines, *Annals of Statistics*, 19, 1-67.

Hornik, Kurt, Maxwell B. Stinchcombe and Halbert White (1989) Multilayer feedforward networks are universal approximators, *Neural Networks*, 2, 359-366.

Mincer, Jacob (1974) *Schooling, Experience and Earnings*, Cambridge: National Bureau of Economic Research.

Murphy, Kevin M. and Finis Welch (1990) Empirical age-earnings profiles, *Journal of Labor Economics*, 8, 202-229.