*The Australian Journal of Agricultural and Resource Economics*, 41:4, pp. 471–97

# Productivity gains from farmer education in China[†]

Tin Nguyen and Enjiang Cheng*

Regression results for five Chinese provinces provide stronger empirical support than hitherto for the role of farmer education in productivity gains. However, these results are largely due to the inclusion of two small groups of outliers: one consists of very poor households whose heads have no education and the other, relatively well-off households whose heads have only three years of education. This article illustrates the need for a cautious interpretation of the regression results of earlier studies, because they could be affected by problems of outliers, multicollinearity and omitted variable bias.

## 1. Introduction

There has long been a considerable interest in the economic effect of *formal* farmer education because agriculture is by size a very important sector in most developing countries. There is general agreement that a large increase in agricultural productivity requires a modernising environment (with more complex technology), in which farmer education becomes more important because of the greater advantages the more educated farmers have over the less educated ones. It is widely believed that agricultural development is just as important as industrialisation in any development strategy and that the key to agricultural development is a rapid switch to a modernising environment, in which special emphasis is placed on high-yielding varieties of seeds, new inputs such as chemical fertilisers and pesticides, and irrigated water. Many economists believe that farmers' education should have a more

[†] We wish to thank the referees of this journal for their thoughtful, meticulous and appropriate comments and suggestions on an earlier draft of this article which have helped to improve its presentation and arguments considerably. We also wish to thank the Australian Centre for International Agricultural Research for their financial support, Chris Findlay and Sue Richardson (University of Adelaide) for their valuable suggestions and encouragement and the participants at the Global Agricultural Science Policy for the Twenty-first Century Conference at Melbourne, 26–28 August 1996, for their helpful comments.

* Tin Nguyen and Enjiang Cheng, Chinese Economy Research Centre, Department of Economics, University of Adelaide, Australia.

important role to play in modern agriculture than in traditional agriculture. For example, Schultz (1975) hypothesised that education is more effective in a modernising environment, while Mellor (1976) considered farmer education in rural areas to be a central ingredient in a strategy to improve agricultural productivity. The criteria for identifying an environment as non-modern include primitive technology, traditional farming practices and crops and little reported innovation or exposure to new methods. The criteria for identifying an environment as modern, conversely, include the availability of new crop varieties, innovative planting methods, erosion control, and the availability of capital inputs such as insecticides, fertilisers, and tractors or machines. Some other indicators of a modern environment are market-oriented production and exposure to extension services (Lockheed, Jamison and Lau 1980, pp. 55–6).

The contribution of farmer education to farm efficiency appears to have received widespread empirical support from the studies surveyed by Lockheed *et al.* (1980) and Phillips (1994). Lockheed *et al.* (1980), in a critical survey of eighteen studies (using 37 sets of farm data), found education to have a positive and statistically significant effect on farm efficiency in 31 data sets and a negative but statistically insignificant effect in the other six data sets. After pooling all the results of the studies included in their survey, Lockheed *et al.* concluded that farm productivity increases, on average, by 7 per cent (10 per cent and 1 per cent in the modern and non-modern environments, respectively) as a result of a farmer's completing four additional years of education.[1] Carrying out a meta-analysis[2] for the relationship between farmer education and farm productivity, covering 30 studies and 59 data sets (which include the studies and data sets reviewed by Lockheed *et al.*), Phillips (1994) found that the weighted gain of four years of schooling is 6 per cent and that the average gain in productivity of farmers in a modernising environment exceeds that of farmers in traditional surroundings (7 per cent and 3 per cent respectively).

[1] The 7.4 per cent is a weighted average of values from those studies for which an estimate could be computed. Phillips (1994) believed that 7.4 per cent was a mistake, as his calculation based on Lockheed *et al.*'s figures and assumptions produces only 5.7 per cent (see Phillips 1994, footnote 1, p. 164). It is not entirely clear whose education it is. See footnote 17 in Phillips (1994) for a further comment on this.

[2] Meta-analysis is a method of analysing results across empirical studies, which has been used more extensively in psychology, education, and the health sciences than in economics. It is an exercise in which the data points are the individual studies as opposed to individual subjects or observations. To adjust for differences in study reliability, Phillips also followed Lockheed *et al.* to weigh the percentage gain of four years' schooling by the reciprocals of their estimated standard error (see Phillips 1994, p. 149 and p. 155).

Note that the above results for the gain attributed to four years of schooling are for a typical farmer (assumed to be the household head) with an average number of years of schooling and not for an illiterate farmer, with zero education. The gain per year for a farmer starting from zero education may be higher than that for a farmer with an average education but the studies surveyed by Phillips and Lockheed *et al*. failed to address this issue.

We believe that farm efficiency depends primarily on the education of the household head, who makes almost all the farm decisions, rather than on the average education of the other working members of the household. To make good decisions even in the case of improved seeds in a modernising environment, it seems to be only necessary to know how to read and write. Hence, in this article we shall investigate the hypothesis that it is literacy (or the first three years of pre-junior primary education) of the household head (rather than his higher education or the average education of farm workers' generally) that makes a significant difference to farm income. For convenience, let this hypothesis be called the *household head literacy* hypothesis. Much empirical literature investigates the effect on farm efficiency of the education of an average farmer with an average education. As far as we are aware, the household head literacy hypothesis has not formally been investigated. Note that in rural China, household heads are usually men who make decisions for the family. Rural China is still very much a male-dominated society. The major duties for women are housework, child-rearing and farm work. This is perhaps the main reason for Chinese parents to prefer boys to girls.

We also believe that most earlier empirical studies have been preoccupied with regression estimates and their statistical significance. But statistical significance can be a misleading indicator of empirical support for a hypothesis concerning the effect of a variable if any one or more of the following conditions hold: (i) there are important outliers; (ii) the sample size is very large (like the one in this article);[3] (iii) important variables are omitted; and (iv) the variable in question is highly correlated with one or more of other regressors. It is shown in most econometrics textbooks that the estimate of the effect of a variable can be (i) imprecise (or unreliable), if this variable does not vary sufficiently independently of other variables

[3] This is because with very large sample sizes, almost any null hypothesis will be rejected (see Gujarati 1995, p. 134; Maddala 1992, p. 32; Kennedy 1985, p. 62). Kalbfleisch and Sprott (1976, pp. 259–72) argue that it is a gross simplification to regard a test of significance as a decision rule for accepting or rejecting a hypothesis and maintain that the purpose of a significance test is just to quantify the strength of evidence in the data against a hypothesis expressed in a (0, 1) scale, not to suggest an accept–reject rule. There are some statisticians who advocate that the significance level used should be adjusted downward as the sample size increases (Lindley 1957, pp. 187–92; Leamer 1978, pp. 88–9 and 104–5).

included in the model, and (ii) seriously biased if it does not vary sufficiently independently of the relevant variables that are omitted from the model. These are the so-called problems of *multicollinearity* and *omitted variable bias*. In view of the prevalence of these problems, researchers can often find significant *t*-ratios for irrelevant variables and hence empirical support for a non-existent effect, and vice versa.[4] Therefore, the magnitudes, signs and significance of the regression coefficients may be highly sensitive to specification errors and the non-fulfilment of the assumptions underlying the estimated model. To deal with this problem, a fragility analysis may be used to examine the range of inferences resulting from the range of believable model specifications (Kennedy 1985, p. 62).

The main aim of this article is to investigate whether education has an effect on farm efficiency and, if so, in what form, using the data from a large sample of 978 households living in five Chinese provinces (Guangdong, Jilin, Jiangxi, Sichuan and Shangdong) in 1995. Our regression results indicate that it is the education of the household head – rather than the average education of farm workers or the age (or experience) of household head or farm workers – that has a significant effect on farm productivity (as proxied by income per farm worker). However, further analysis reveals that only the first three years of the household head's education (for literacy) are important in terms of their effect on farm efficiency and his education beyond this has little effect. It should be pointed out that a household head with three years of education may not been regarded as literate. Obviously it depends on the quality of the education received by him (e.g. qualifications and experience of the teaching staff). Under normal conditions in China, three years' education would enable one to read the newspapers and instructions for the use of chemical fertilisers, pesticides and other farm inputs. Still closer examination throws doubt on the benefit of literacy itself. The highly statistically significant estimate we obtain for the effect on farm income of the first three years of education of a household head may simply reflect the existence of 48 very poor households with illiterate heads. It is possible that these households are poor for reasons other than their illiteracy and, if so, its removal would not lead to the expected dramatic increase in income suggested by our regressions. It is also possible that illiteracy is the

[4] That is, they find no support for relevant variables, or effects which do in fact exist. In so far as a regressor is highly correlated with the omitted relevant variables, the risk of wrongly rejecting the true null hypothesis (that the regressor has no effect) may well be considerably greater than the chosen level of significance. Moreover, it is misleading to focus attention on the risk of type I error (i.e. the level of significance) and overlook the importance of type II error (of wrongly accepting a false null hypothesis). The risk of type II error will also be affected by a higher correlation between each regressor and the omitted variables.

result rather than the cause of poverty.[5] However, as elaborated in the concluding section, the gains in terms of income or productivity are only one element of a complete cost–benefit analysis of an education program which encompasses many other private and social benefits from formal education.

The following section discusses: (i) the rationale for choosing 'farm income' as the measure of productivity gains from farmer education; (ii) the possible mis-specification in using aggregate data; and (iii) the strategy for choosing models. In the next three sections, the various models used to produce the results are described, the nature and sources of the survey data used are discussed and the results are presented and evaluated. The final section summarises the main findings and discusses their implications for policy and further research.

## 2. Conceptual, theoretical and methodological considerations

It is well known that the econometric results are often sensitive to the choice of (i) the dependent variable; (ii) the level of disaggregation for the data; and (iii) the strategy for model selection. This section will briefly discuss the reasons for our choice concerning (i)–(iii) above.

### 2.1 The dependent variable

According to the survey by Lockheed *et al.* (1980, p. 45), analysis of 23 of the 37 data sets used the value of crop production as the dependent variable, although most of these studies were described as studies of production. The choice of a dependent variable depends crucially on the definition of what each household is assumed to maximise. A reasonable objective for households to maximise appears to be total household income, which consists of

[5] Logically, causality can be refuted but not confirmed, using any test, statistical or otherwise. If the effects on $y$ of *all* the variables other than $x$ can be removed or neutralised when $x$ varies, via a controlled experiment or regression, so that the effect of $x$ on $y$ can be isolated, then a single such observation will be sufficient to confirm that $x$ causes $y$. The problem is that we can never be sure that we can control all possible influences on $y$ which may covary with $x$. The Granger causality test is not really a test of causality but a test of precedence, i.e. whether $x$ 'precedes' $y$, since that $x$ occurs before $y$ is necessary but not sufficient for $x$ to cause $y$. In purely logical terms, as Karl Popper recognised, there cannot be a completely causal science; the real role of observation and experiment is to refute theories, not to confirm them. Hume went further to say that no matter how many times an experiment agrees with a hypothesis, it does not make the hypothesis more likely to be true (Mitchell and Stein 1990, p. 1436; Crofton 1990, pp. 58–9). Econometricians should follow scientists and accept the view of the nineteenth-century British philosopher, Francis Bacon, that 'an agreement of a conclusion with an actual observation does not itself prove the correctness of the hypothesis from which the conclusion is derived. It merely renders the premise that much more plausible.'

farm income and off-farm income. While the former consists of earnings from the sales of crops, animal and aquatic products (at farm-gate prices), the latter consists of earnings from off-farm activities and wages earned by household members working elsewhere. Although off-farm income has been growing since the economic reform launched in 1979, farm income still represents the major source of income for most Chinese farmers. Since the household heads have greater influence over farm decision-making than off-farm decision-making and since the focus of this study is on the contribution of their education to farm efficiency, farm income is a more appropriate dependent variable than off-farm income or total income.[6]

Average farm income per worker is used as the dependent variable instead of total farm income for the following three reasons: (i) to provide a convenient test of returns to scale (i.e. the $t$-test on coefficient of the number of workers) and to facilitate the imposition of constant returns to scale as a restriction on the model (by omitting the number of workers from the regression); (ii) to reduce heteroscedasticity which tends to occur when the variables are not scaled; and (iii) to avoid producing a spuriously large $R^2$ which might inspire an exaggerated confidence in the 'overall performance' of the model. Note that $R^2$ tends to be larger for an equation involving unscaled variables because most of the inputs tend to increase with the number of workers.

## 2.2 Regression with aggregate data

For ease of exposition, let us assume without any loss of generality that the households produce $m$ crops and nothing else. Farm income ($Q$) may be defined as follows:

$$Q = \sum_{i}^{m} \Pi_i X_i = \sum_{i}^{m} \Pi_i \{G_i(S) F_i(I_i)\} = \Phi(S; \Pi_1, \ldots, \Pi_m; I_1, \ldots, I_m) \qquad (1)$$

where $\Pi_i$ and $X_i$ $\{= G_i(S)F_i(I_i)\}$ are respectively the price and output of crop $i$; $G_i(S)$ is a scale variable representing total factor productivity; $S$ is a vector of household characteristics such as education and experience; and $I_i$ is a vector of inputs used in the production of crop $i$. Equation 1 requires information on individual crops which are usually unavailable. Therefore, to proceed with the data analysis, econometricians often have no choice but to

[6] In fact, both off-farm income and total income have been tried as the dependent variables but the results are poorer both in terms of statistical significance and plausible magnitudes. This is largely due to the fact that, unlike farm income, off-farm income depends on many variables on which data are not readily available.

assume that 1 can be approximated by a function which involves only aggregate data, e.g.:

$$Q = \Psi(S; I) \tag{2}$$

where $I = \sum_i^m I_i$ is a vector of the totals of inputs for all crops. By assuming that the price of each crop is the same for all households, the price variables are absorbed into the intercept (not shown in 2). Mathematically, it is impossible to derive (2) from (1) without making highly implausible simplifying assumptions about the functional forms of $G_i$ and $F_i$, assumptions which would necessarily reduce the generality of the results. If (1) is true, then the parameters of (2) cannot be constant. While recognising the inadequacy of (2) as an approximation for (1), we follow other studies to estimate (2), because reliable data on inputs are available only in aggregate form. Note that even if disaggregated data on inputs and crops are available, estimation of (1) is likely to encounter the problem of multicollinearity as the same inputs allocated to different crops are likely to be highly correlated. Furthermore, the model will contain a large number of parameters, making it even harder to interpret the results.

Economic theory tells us little more about the functional form of (1) or (2) than that the first and second derivatives of $Q$ with respect to each input are positive and negative respectively, i.e. the marginal productivity of each input is positive and diminishing. This theoretical requirement is consistent with several functional forms (e.g. double-log, trans-log, quadratic) to represent the relationship between $Q$ and $I$. Our approach here is to start from a general functional form, e.g. with squares and cross-products of variables and their logarithms, and progress to a simpler form, e.g. by deleting successively those regressors which have the wrong signs or are insignificant at 5 per cent.

### 2.3 Model selection

Although this topic is of special importance for all empirical works, it will be discussed only briefly here, because it is well documented elsewhere (e.g. Maddala 1992, pp. 490–506; Gujarati 1995, pp. 452–99). According to the conventional approach to econometric modelling, the statistical model underlying the data is assumed to be known at the outset and the problem is simply one of obtaining good estimates of the parameters in the model. In reality, however, the choice of a model is almost always made after some preliminary data analysis (Maddala 1992, p. 490). For example, we may start with a regression model with only a few plausible variables and proceed to add 'new' variables using some diagnostic tests or search procedures to 'build up' the model. This is the so-called *bottom-up* approach. The alternative approach to this is the *top-down* or *general-to-specific* approach, advocated and refined by Hendry and popularised by Hendry and Richard and Gilbert (Hendry and Richard 1983; Gilbert 1986). According to the latter approach, we start with a very general model, with a specification of the functional form and lag structure which is typically more complicated than deemed necessary and progressively simplify it with a sequence of 'simplification tests'. Maddala (1992, p. 494) aptly summarises the traditional average economic regression approach and the new general-to-specific approach as: 'excessive presimplification with inadequate diagnostic testing' and 'intended overparametrisation with data-based simplification' respectively.[7]

In this study, following the general-to-specific approach, we start with a model as general as our data set permits. The initial model consists of (i) all the variables which appear relevant and on which we have data; (ii) their squares, cross-products and some other appropriate transformations (e.g. logarithms); and (iii) a large number of dummy variables for slopes and intercepts. We then estimate this model and simplify it by successively removing all the variables and their squares and cross-products (i) which are statistically insignificant at the 10 per cent level; (ii) which do not have the expected signs; (iii) which have implausible magnitudes or (iv) which cause the coefficients of other variables, known to be important, to have wrong signs and magnitudes. In the next section, instead of specifying a general model to start from, we specify a model that is similar to those specified in many previous studies and that happens to be quite close to the final model reached after a series of simplifying tests.

## 3. Model specification

It is useful to present first the simplest model and then to introduce new regressors to obtain successively more and more general models. Testing and estimating these models using aggregate cross-section data will necessarily encounter various econometric problems (see Griliches 1977, for an excellent discussion of these problems).

[7] One advantage of the general-to-specific approach is that, if a true model or data-generating-mechanism (DGM) exists, then following this approach different econometricians are likely to arrive eventually at similar models (which may differ little from the DGM). The same econometricians, adopting the conventional approach, may arrive at substantially different models, because they each may start from a very different simple model and the final models are often sensitive to the choice of the starting ones. Since their models are widely different, only a few or none of their models may get close to the DGM.

### 3.1 The models

Consider the following popular form for the regression model (Lockheed *et al.* 1980, p. 40; Phillips 1994, p. 150):

$$L_Q = \sum_m \kappa_m P_m + \sum_i \alpha_i L_i + \sum_j \beta_j Z_j + \sum_j \gamma_j Z_j^2 + u \tag{3}$$

$$(1 > \alpha_i > 0 \text{ for all } i;\ \beta_j > 0 \text{ and } \gamma_j < 0 \text{ for all } j.)$$

where

$L_Q$ = Logarithm of $Q$
$P_m$ = 'Intercept' dummy variable for province $m$, where $m$ = Guangdong, Jilin, Jiangxi, Sichuan, Shangdong ($P_m = 1$ when household in province $m$ and $P_m = 0$, otherwise).
$L_i$ = Logarithm of $i$ where $i = ST, CF, LD, CT, T, K$.
$Z_j$ = Level of $j$, where $j = SH, AH, SM, AM$; however, $Z_j$ is replaced by $j$ in all the Tables of Results (e.g. $Z_{SH}$ and $Z_{AH}$ are replaced by $SH$ and $AH$ respectively).

and

$Q$ = total farm income in yuan (from crops and animal products)
$ST$ = sown areas in mu (1 hectare = 15 mu)
$CF$ = chemical fertilisers in yuan
$LD$ = labour days (1 semi-labour day = 0.7 full labour day)
$CT$ = other inputs in yuan (total value of inputs other than labour, fertiliser and seeds)
$K$ = total value of capital stock owned by a household valued at current price in yuan ($Q$, $ST$, $CF$, $LD$, $CT$ and $K$ are *all* per worker)
$SH$ = number of schooling years of the household head
$AH$ = the age of the household head
$SM$ = average number of schooling years of household workers except the head
$AM$ = average age of the household workers except the head
$T$ = number of household workers
$Z_{SH_1}$ = a slope dummy variable defined as followed: $Z_{SH_1} = Z_{SH}$ if the household head has 1–3 years of education and $Z_{SH_1} = 0$, otherwise.

Equation 1 postulates a double-log relationship between $Q$ and each of the input variables (represented by the $L_i$s), i.e. $ST$, $CF$, $LD$, $CT$, $T$ and $K$; and a quadratic semi-log relationship[8] between $Q$ and each of the education and age variables (represented by the $Z_j$s), i.e. $SH, AH, SM$ and $AM$. The marginal productivity of each input is positive but diminishing as dictated by economic theory, since $(1 > \alpha_i > 0)$. The marginal effect of a year of education or age can be increasing initially and then diminishing after a point.[9] The coefficient of $LT$ (i.e. the logarithm of the number of workers) is a measure of non-constant returns to scale. Returns to scale are increasing or decreasing if this coefficient is positive or negative. To allow (and test) for non-constant returns to scale, $LT$ is included in the estimated equation, its coefficient is expected to be negative for diminishing returns to scale. Note that all the input and output variables are in terms of per worker.

To test the hypothesis that the first few years of schooling would have a stronger effect on farm income than subsequent years of education, the expression $\delta Z_{SH_1}$ is added to equation 1 to give:

$$L_Q = \sum_m \kappa_m P_m + \sum_i \alpha_i L_i + \sum_j \beta_j Z_j + \sum_j \gamma_j Z_j^2 + \delta Z_{SH_1} + u \tag{4}$$

Note that $\delta > 0$ represents the increase in the coefficient of the household head's education associated with the first three years of education. To test the hypothesis that the households whose heads have one to three years of education have a different set of output elasticities, we can add the following expression $\sum_i \alpha_{i0} L_{i0}$ to equation 4:

$$L_Q = \sum_m \kappa_m P_m + \sum_i \alpha_i L_i + \sum_i \alpha_{i0} L_{i0} + \sum_j \beta_j Z_j + \sum_j \gamma_j Z_j^2 + \delta Z_{SH_1} + u \tag{5}$$

[8] Since education is usually measured in years of schooling and includes a value of zero for farmers with no education whatsoever, a double-log relationship between education and income cannot be postulated. On the other hand, the double-log relationship between income and farmers' age to represent experience is of course possible. In fact, we have experimented with a translog model involving all variables other than education and found they perform poorer on the basis of econometric criteria than the model represented by (1), e.g. the coefficients of the squares and cross-products of the logarithms of the variables are all statistically non-significant.

[9] For example, the marginal effect of $SH$ is:

$$\frac{\partial \mathrm{Log}(Q)}{\partial SH} = \left(\frac{1}{Q}\right) \frac{\partial Q}{\partial SH} = \beta_{SH} + 2\gamma_{SH} SH; \text{ hence: } \frac{\partial Q}{\partial SH} = (\beta_{SH+2\gamma_{SH}} SH)Q, \text{ with slope:}$$

$$\frac{\partial^2 Q}{\partial SH^2} = (\beta_{SH} + 2\gamma_{SH} SH)\,\frac{\partial Q}{\partial SH} + 2\gamma_{SH} Q = \{(\beta_{SH+2\gamma_{SH}} SH)^2 + 2\gamma_{SH}\}Q$$

Since $\gamma_{SH} < 0$, $\dfrac{\partial^2 Q}{\partial SH^2}$ becomes negative, as $SH$ increases beyond a point, $\dfrac{\partial Q}{\partial SH}$ also becomes negative.

where $L_{i0}$ is a slope dummy variable defined as follows: $L_{i0} = L_i$ if the household head has *zero* education and $L_{i0} = 0$, otherwise; where $i = ST$, $CF, LD, CT, T, K$ (e.g. $L_{CF0}$ and $L_{LD0}$ denote the 'slope' dummy variables for chemical fertilisers and labour days respectively) and $\alpha_{i0}$ represents the difference between the coefficient of $L_i$ for households whose heads had zero education and that for all other households. In other words, the coefficient of $Z_i$ is $\alpha_i + \alpha_{i0}$ for households whose heads have zero education and $\alpha_i$ for others.

## 4. Data

Data for the present study came from a sample survey of 1 041 rural households conducted in 1995. The survey was designed and conducted jointly by the Chinese Economy Research Unit of the University of Adelaide, Australia and the Ministry of Agriculture in China. The following data are collected for each household: crop production and uses; number; education and age of the household head and other household members; occupation (including official status) of the household head; various direct inputs for grain production, including sown areas and number of plots, labour working days, chemical fertilisers, and other variable inputs; capital stock owned by households. This data set includes detailed information on human capital variables and occupation of household heads rarely available for such a large sample.

The households included in this survey were sampled from five provinces of China: Guangdong, Jilin, Jiangxi, Sichuan and Shangdong. Of 1 041 sample households, 215 were from Guangdong, 201 from Jilin, 205 from Jiangxi, 200 from Sichuan and 220 from Shangdong. Although the selection of sample provinces and counties was deliberately biased towards major grain-producing areas, the sample provinces can be regarded as representative of rural China, except the very poor north-west. The five provinces are located in different parts of China. Guangdong, Shangdong and Jilin provinces are located in the more developed coastal areas of Eastern China, with Guangdong in the rich Pearl Delta down in the South, Jilin in the industrial area of the North-east and Shangdong in north China. Sichuan and Jiangxi are poor inland provinces, with Sichuan located in the south-west and Jiangxi in the middle south. (For a comparison of the basic economic statistics of the five provinces with those of the whole of China, see Wu 1995.) The largest sample used in this article consists of 978 observations, since 63 out of a total of 1 041 households did not provide detailed information on the variables of interest to this study (e.g., education, income, experiences, other farm inputs) and had to be excluded.

## 5. Results

Following the general-to-specific approach to model selection, we first estimate equation 4, with all the variables included, and systematically remove the variables with insignificant coefficients and/or wrong signs. All the equations successively estimated are presented in table 1. It can be seen that all the variables in column 6, in table 1, have the expected signs and, of its twelve coefficients, ten are significant at 0.1 per cent, one at 1 per cent and one at 5 per cent.[10] Columns 1–5 reveal that (i) farmers' experience (as proxied by farmers' age, $AH$ and $AM$); (ii) average education of farm workers ($SM$); and (iii) capital do not have a statistically significant effect on income. It is possible that variation in age over a certain interval may have an effect on income. To test this we include a set of dummy variables ($AH_i$, where $Ah_i = AH$ for households in age group $I = 1, 2, \ldots, m$ and $Ah_i = 0$, otherwise) and fail to obtain significant or meaningful results for alternative age classifications. As the farm worker gets older, he may become more productive because of greater experience, but this is offset by poorer health or weaker physical strength associated with age. Experience may be more important for non-farming activities than for farming activities. The coefficient of (the logarithm of) the number of farm workers ($LT$), though negative as expected, is not statistically significant even at 10 per cent, suggesting that returns to scale are constant. The insignificance of the number of workers is not due to a high correlation between the number of workers and the average number of days worked per worker, because this correlation is in fact not high. As the number of workers increases, the total number of days worked would increase, while the average number of days worked per worker would decline. There would be a high negative correlation between the average number of days worked and the number of workers only if total labour days are fixed or marginal productivity of labour days is zero, which is far from being the case in this study (see table 1). It can be seen from column 7 that the coefficient of the household head's education ($SH$) remains significant even with the omission of $SH1$. There is a small variation in the intercept across the five provinces, ranging from 4.24

[10] As is usual with cross-section data, the null of homoscedasticity cannot be accepted using the White test for most of our equations. With the presence of heteroscedasticity, the OLS estimates are still unbiased but are no longer efficient and their standard errors produced by the usual formulae are biased and inconsistent. Here, the $t$-ratios presented are based on White heteroscedastic-consistent estimates of the standard errors. We do not attempt to improve the efficiency of the estimates by using Generalised Least Squares, because this procedure could produce less efficient results if the exact form of heteroscedasticity is not known.

**Table 1** Regression equation with insignificant variables successively removed (dependent variable: income per worker)

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| *P*1 | 3.0470 | 3.0363 | 3.233 | 3.1993 | 3.8100 | 4.240 | 4.2875 |
| | (5.860) | (5.922) | (7.035) | (7.035) | (13.758) | (24.841) | (24.512) |
| *P*2 | 3.3331 | 3.3214 | 3.5000 | 3.4593 | 4.0689 | 4.4931 | 4.5387 |
| | (6.515) | (6.586) | (7.742) | (7.765) | (14.932) | (26.650) | (26.200) |
| *P*3 | 3.1781 | 3.1664 | 3.3672 | 3.3319 | 3.9388 | 4.3705 | 4.4165 |
| | (6.290) | (6.358) | (7.512) | (7.519) | (14.457) | (26.192) | (25.812) |
| *P*4 | 3.1218 | 3.1102 | 3.2985 | 3.2621 | 3.8739 | 4.2944 | 4.3447 |
| | (6.105) | (6.163) | (7.267) | (7.271) | (14.161) | (25.303) | (24.860) |
| *P*5 | 3.2165 | 3.2048 | 3.3848 | 3.3426 | 3.9584 | 4.3805 | 4.4282 |
| | (6.258) | (6.330) | (7.426) | (7.442) | (14.537) | (26.180) | (25.769) |
| *LST* | 0.2233 | 0.2231 | 0.2742 | 0.2748 | 0.2721 | 0.2707 | 0.2673 |
| | (5.440) | (5.457) | (9.032) | (9.048) | (8.949) | (8.913) | (8.574) |
| *LCF* | 0.3060 | 0.3062 | 0.2895 | 0.2913 | 0.2888 | 0.2907 | 0.2921 |
| | (8.943) | (8.949) | (9.844) | (9.914) | (9.795) | (9.860) | (9.786) |
| *LDF* | 0.1912 | 0.1912 | 0.1962 | 0.2002 | 0.2001 | 0.2003 | 0.2037 |
| | (7.510) | (7.614) | (8.677) | (9.197) | (9.322) | (9.352) | (9.478) |
| *LCT* | 0.0585 | 0.0584 | 0.0525 | 0.0529 | 0.0539 | 0.0593 | 0.0616 |
| | (1.589) | (1.583) | (1.814) | (1.827) | (1.847) | (2.025) | (2.051) |
| *SH* | 0.066 | 0.0654 | 0.0795 | 0.0793 | 0.0849 | 0.0827 | 0.0690 |
| | (4.085) | (4.062) | (4.953) | (4.964) | (5.481) | (5.381) | (4.543) |
| *SH*1 | 0.1031 | 0.1032 | 0.1217 | 0.1208 | 0.1227 | 0.1213 | |
| | (3.602) | (3.646) | (4.451) | (4.449) | (4.495) | (4.457) | |
| *SHSQ* | – 0.0029 | – 0.0029 | – 0.0037 | – 0.0037 | – 0.0041 | – 0.0039 | – 0.0033 |
| | (2.841) | (2.841) | (2.841) | (2.841) | (4.073) | (2.841) | (3.352) |
| *AH* | 0.034 | 0.035 | 0.0284 | 0.0258 | 0.0230 | | |
| | (2.339) | (2.352) | (2.060) | (1.896) | (1.877) | | |
| *AHSQ* | – 0.0004 | – 0.0004 | – 0.0003 | – 0.0003 | – 0.0003 | | |
| | (2.379) | (2.389) | (2.050) | (1.924) | (1.697) | | |
| *AM* | 0.0237 | 0.0238 | 0.0232 | 0.0245 | | | |
| | (1.585) | (1.591) | (1.884) | (2.009) | | | |
| *AMSQ* | – 0.0003 | – 0.0003 | – 0.0003 | – 0.0003 | | | |
| | (1.365) | (1.378) | (1.697) | (1.765) | | | |
| *LT* | – 0.0730 | – 0.0735 | – 0.0454 | | | | |
| | (1.145) | (1.161) | (0.815) | | | | |
| *LK* | 0.0134 | 0.0136 | | | | | |
| | (1.013) | (1.029) | | | | | |
| *SM* | – 0.0015 | | | | | | |
| | (0.147) | | | | | | |
| *SMSQ* | 0.0001 | | | | | | |
| | (0.531) | | | | | | |
| *ADJ-R*2 | 0.5668 | 0.5678 | 0.5812 | 0.5814 | 0.5807 | 0.5801 | 0.5729 |
| *F-STATS* | 59.89 | 67.07 | 85.49 | 91.18 | 104.75 | 123.73 | 132.03 |
| *N* | 856 | 856 | 975 | 975 | 978 | 978 | 978 |

Notes:
[a] The figure in brackets under each coefficient is its *t*-ratio (in absolute value), which is based on the heteroscedastic-consistent estimate of its standard error.
[b] For definitions of all symbols, see p 479.

(Guangdong) to 4.49 (Jilin), representing the effect of the inter-provincial variation in climatic and topographical characteristics.

Column 6 is estimated for each of the five provinces and the results are presented in table 2. The coefficients of all the variables have the right signs in each province equation but are all less significant individually than in the pooled equation (for the total of all five provinces). This is partly due to the larger number of degrees of freedom in the pooled equation. The better results for pooled sample for education do suggest that the inter-provincial covariation of education and income is also responsible for the significant partial relationship between education and income for households in the pooled sample. The magnitude of the coefficient of the household head's education also varies considerably across the provinces. It is largest for Sichuan (0.12) and smallest for Guangdong (0.03). The coefficient of $SH1$ is also largest for Sichuan (0.19) and smallest for Guangdong (0.04).

**Table 2** Regression equations by provinces (dependent variable: income per worker)

| | Guangdong | Jilin | Jiangxi | Sichuan | Shangdong | Total |
|---|---|---|---|---|---|---|
| $C$ | 4.738 | 3.669 | 4.487 | 4.059 | 4.431 | N/A |
| | (12.939) | (10.611) | (11.488) | (10.217) | (11.097) | N/A |
| $LST$ | 0.350 | 0.137 | 0.318 | 0.330 | 0.216 | 0.271 |
| | (5.517) | (1.964) | (5.058) | (4.906) | (3.005) | (8.913) |
| $LCF$ | 0.294 | 0.386 | 0.386 | 0.208 | 0.212 | 0.291 |
| | (3.916) | (4.760) | (7.937) | (3.183) | (3.418) | (9.860) |
| $LDF$ | 0.134 | 0.186 | 0.099 | 0.304 | 0.284 | 0.201 |
| | (3.157) | (4.990) | (1.997) | (5.401) | (4.193) | (9.352) |
| $LCT$ | 0.055 | 0.144 | 0.057 | 0.041 | 0.083 | 0.059 |
| | (0.693) | (2.060) | (0.932) | (0.748) | (1.261) | (2.025) |
| $SH$ | 0.026 | 0.103 | 0.055 | 0.119 | 0.073 | 0.083 |
| | (0.989) | (2.538) | (1.291) | (3.489) | (2.534) | (5.381) |
| $SH1$ | 0.044 | 0.086 | 0.123 | 0.190 | 0.078 | 0.121 |
| | (0.713) | (1.413) | (2.208) | (3.590) | (1.308) | (4.457) |
| $SHSQ$ | – 0.001 | – 0.005 | – 0.003 | – 0.005 | – 0.004 | – 0.004 |
| | (0.599) | (2.172) | (0.786) | (1.944) | (2.180) | (3.902) |
| $ADJ\text{-}R2$ | 0.533 | 0.694 | 0.431 | 0.551 | 0.393 | 0.580 |
| $F\text{-}STATS$ | 34.441 | 61.384 | 22.197 | 35.060 | 18.733 | 123.726 |
| $N$ | 206 | 187 | 197 | 195 | 193 | 978 |

Notes:

[a] This total is the same as column 6 in table 1; it includes five intercepts, one for each county, in the place of a single intercept. The values of these intercepts are included in table 2.

[b] The figure in brackets under each coefficient is its $t$-ratio (in absolute value), which is based on the heteroscedastic-consistent estimate of its standard error.

[c] For definitions of all symbols, see p. 479.

[d] The 5 per cent, 1 per cent and 0.1 per cent critical values for $t$-ratio are approximately 1.66, 2.36 and 3.16 (one-tailed test) and 1.98, 2.58 and 3.30 (two-tailed test) respectively, for number of degrees of freedom equal to 120 or larger.

**Table 3** Implications of the regression results for the economic effects of education

| Education ($SH$) in years | Residual income index base: $SH=0$ % | Residual income index base $SH=5$ % | Change from base income $SH=0$ % | Change from base income $SH=5$ % |
|---|---|---|---|---|
| 0 | 100.00 | 72.84 | 0.00 | −27.16 |
| 1 | 122.15 | 88.98 | 22.15 | −11.02 |
| 2 | 148.07 | 107.85 | 48.07 | 7.85 |
| 3 | 178.10 | 129.73 | 78.10 | 29.73 |
| 4 | 130.86 | 95.32 | 30.86 | −4.68 |
| 5 | 137.29 | 100.00 | 37.29 | 0.00 |
| 6 | 142.92 | 104.10 | 42.92 | 4.10 |
| 7 | 147.64 | 107.54 | 47.64 | 7.54 |
| 8 | 151.35 | 110.24 | 51.35 | 10.24 |
| 9 | 153.95 | 112.14 | 53.95 | 12.14 |
| 10 | 155.39 | 113.19 | 55.39 | 13.19 |
| 11 | 155.65 | 113.37 | 55.65 | 13.37 |
| 12 | 154.70 | 112.68 | 54.70 | 12.68 |
| 13 | 152.58 | 111.14 | 52.58 | 11.14 |
| 14 | 149.33 | 108.77 | 49.33 | 8.77 |
| 15 | 145.02 | 105.63 | 45.02 | 5.63 |

Notes:
[a] Residual income is the income after the effects of the variation of inputs being removed.
[b] Let $a$, $b$ and $c$ be respectively the coefficients of $SH$, $SHSQ$ and $SH1$ of column 6, table 1, (i.e. 0.008268, 0.12129 and −0.00386 respectively) then the index for residual income is:

$$Q = A\{\exp[(a + c + b^{*}SH)^{*}SH]\} \qquad \text{for } SH = 0, 1, 2, 3$$

$$Q = A\{\exp[(a + b^{*}SH)^{*}SH]\} \qquad \text{for } SH > 3$$

where $A$ can be taken to be equal to the average income per worker for group $G0$ for the Total of all five provinces (i.e. $A = 1358.3$ yuan), given in table 5. It can be seen that the marginal effect of education is negative between the third year and the fourth year of education and becomes consistently negative after the eleventh year. The first three years of education will raise income per worker by 78 per cent of base income ($SH = 0$). Four additional years of education from base income ($SH = 5$) yields about 12 per cent of this base income.

The second column of table 3 gives an index of the residual income per worker (after the effects of differences in the levels of inputs per worker and provincial characteristics have been removed) with the value of the index equal to 100 for the household whose head has zero education. The third column of table 3 gives the same information as the second column, except that a value of 100 is given for the household whose head has five years of education. The fourth column gives the percentage change in income per worker resulting from one, two, three, . . . , fifteen extra years of schooling, starting from zero education. (For example, the income per worker would have increased by 22 per cent, 48 per cent and 78 per cent if its head had one, two, and three years of education respectively instead of zero education.) The fifth column gives the same information as the fourth

**Table 4** Regression equations: pooled data for five provinces (dependent variable: income per worker)

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| *P*1 | 4.240 | 4.287 | 4.312 | 4.516 |
| | (24.841) | (24.512) | (20.531) | (22.578) |
| *P*2 | 4.493 | 4.539 | 4.558 | 4.761 |
| | (26.650) | (26.200) | (22.211) | (24.184) |
| *P*3 | 4.370 | 4.417 | 4.441 | 4.644 |
| | (26.192) | (25.812) | (21.620) | (23.677) |
| *P*4 | 4.294 | 4.345 | 4.371 | 4.572 |
| | (25.303) | (24.860) | (21.015) | (22.953) |
| *P*5 | 4.381 | 4.428 | 4.452 | 4.654 |
| | (26.180) | (25.769) | (21.824) | (23.819) |
| *LST* | 0.271 | 0.267 | 0.281 | 0.283 |
| | (8.913) | (8.574) | (9.060) | (9.033) |
| *LCF* | 0.291 | 0.292 | 0.306 | 0.311 |
| | (9.860) | (9.786) | (10.359) | (10.517) |
| $LCF_0$ | | | – 0.208 | – 0.228 |
| | | | (3.341) | (3.668) |
| *LDF* | 0.200 | 0.204 | 0.188 | 0.190 |
| | (9.352) | (9.478) | (8.640) | (8.662) |
| $LDF_0$ | | | 0.218 | 0.190 |
| | | | (3.209) | (2.787) |
| *LCT* | 0.059 | 0.062 | 0.051 | 0.049 |
| | (2.025) | (2.051) | (1.716) | (1.650) |
| *SH* | 0.083 | 0.069 | 0.068 | 0.017 |
| | (5.381) | (4.543) | (2.335) | (0.693) |
| *SH*1 | 0.121 | | 0.113 | |
| | (4.457) | | (3.394) | |
| *SHSQ* | – 0.004 | – 0.003 | – 0.003 | – 0.0004 |
| | (2.841) | (3.352) | (1.907) | (0.279) |
| *ADJ-R*2 | 0.580 | 0.573 | 0.584 | 0.579 |
| *F-STATS* | 123.73 | 132.025 | 106.390 | 113.004 |
| *N* | 978 | 978 | 978 | 978 |

Notes:
[a] The figure in brackets under each coefficient is its *t*-ratio (in absolute value), which is based on the heteroscedastic-consistent estimate of its standard error.
[b] For definitions of all symbols, see p. 479.
[c] The coefficients of the two dummy variables $LCF_0$ and $LDF_0$ respectively represent the shifts in the output elasticities of fertilisers and workers' days of working associated with the households with illiterate heads, i.e. having zero education. Note that $LCF_0 = LCF$ for households with illiterate heads and $LCF_0 = 0$, otherwise.

column, except that the change in income per worker associated with the extra years of education is expressed as a percentage of the income of the household whose head has initially five years of education rather than zero education. (For example, the income per worker of a household with a head having five years of education would have increased by about 4 per cent, 7 per cent, 10 per cent and 12 per cent had its head had one, two, three and four additional years of education respectively.)

The effects of four additional years of education[11] for a household, whose head had initially five years of education (i.e. two years above the average number of years of education for all household heads) were estimated at 10 per cent and 7 per cent by Lockheed *et al.* and Phillips respectively for the modern environment. From the fifth column of table 3, it can seen that the effect of four years of education for the same household would be 12 per cent. This is equal to the percentage increase in income per worker associated with an increase in the household head's education from five to nine years (see column 5 of table 3, note that in our sample, the mean education for all household heads is about seven and a half years, which is almost the same as the value in Lockheed *et al.* and Phillips). Our figure for the effect of the education of the household head with approximately an average education appears to be rather high by comparison. Our figure of 78 per cent for the effect of the first three years of education for the household head is surprisingly large – too large to be believable.

To help explain this very large effect of household head's literacy on household income, an equation which includes the same variables as column (6) but without $SH$, $SH1$ and $SHSQ$ (to denote $SH^2$) is re-estimated for each region as well as total. The mean of the residuals (in antilog form) for each education grade minus unity is given in the second column of table 5 and plotted in figure 2. This represents, for each education group, the percentage deviation of the actual income per worker from its conditional expectation (given the values of land, labour days, fertilisers, other inputs, etc.). For each region as well as for total of all five, the distribution of the number of households by household head's education grade is shown graphically in figure 1. The definition of each grade in terms of years of schooling is given in the fourth column of table 5.

It can be seen from table 5 that, for the five provinces as a whole, there are 48 households (in education group $G0$) whose heads had zero education and 43 households (in education group $G1$) whose heads had one to three years of education. From figure 2 and table 5 (column 2), it can be seen that households in these two groups (i.e. $G0$ and $G1$) for the total of all five provinces had respectively income per worker 31 per cent below and 15 per cent above the average for all education grades, after the effects of all inputs

[11] Lockheed *et al.* (1980, p. 40), define education (E) as the education of household head (in years of formal education completed). The word education is later used without any qualification to say whose education it is. Phillips (1994) does not indicate anywhere whether education refers to the average education of the household workers or the education of the household head.

**Table 5** Residual income, household head's education and other variables: by education grade[a]

| | $N$ | $N$(%) | $YF$ | $YFR$ | $SH$ | $EG$ |
|---|---|---|---|---|---|---|
| Guangdong | | | | | | |
| | 7 | 3.4 | 1157.5 | −12.5 | 0.0 | $G0$ |
| | 7 | 3.4 | 2056.2 | −1.0 | 2.6 | $G1$ |
| | 22 | 10.7 | 1385.1 | −0.3 | 4.3 | $G2$ |
| | 68 | 33.0 | 1477.8 | −1.9 | 6.7 | $G3$ |
| | 68 | 33.0 | 1559.9 | 2.4 | 9.2 | $G4$ |
| | 24 | 11.7 | 1534.9 | 1.2 | 12.0 | $G5$ |
| | 10 | 4.9 | 1318.0 | 4.8 | 14.5 | $G6$ |
| Jilin | | | | | | |
| | 11 | 5.9 | 1410.5 | −34.2 | 0.0 | $G0$ |
| | 7 | 3.7 | 2677.9 | 8.4 | 2.6 | $G1$ |
| | 19 | 10.2 | 2437.2 | −16.7 | 4.3 | $G2$ |
| | 61 | 32.6 | 3032.7 | 14.7 | 6.5 | $G3$ |
| | 65 | 34.8 | 2923.4 | −2.4 | 9.2 | $G4$ |
| | 19 | 10.2 | 3504.1 | 4.6 | 12.0 | $G5$ |
| | 5 | 2.7 | 2600.3 | −2.6 | 14.0 | $G6$ |
| Jiangxi | | | | | | |
| | 10 | 5.1 | 1358.5 | −19.6 | 0.0 | $G0$ |
| | 9 | 4.6 | 2221.0 | 13.1 | 2.4 | $G1$ |
| | 36 | 18.3 | 1580.8 | −6.1 | 4.4 | $G2$ |
| | 56 | 28.4 | 1832.3 | 3.7 | 6.4 | $G3$ |
| | 68 | 34.5 | 1911.6 | 0.5 | 9.1 | $G4$ |
| | 16 | 8.1 | 2081.2 | 5.1 | 12.0 | $G5$ |
| | 2 | 1.0 | 1668.9 | 7.3 | 14.5 | $G6$ |
| Sichuan | | | | | | |
| | 9 | 4.6 | 1506.9 | −43.2 | 0.0 | $G0$ |
| | 11 | 5.6 | 2200.9 | 34.5 | 2.8 | $G1$ |
| | 45 | 23.1 | 1352.2 | −17.0 | 4.1 | $G2$ |
| | 52 | 26.7 | 2059.1 | 2.7 | 6.4 | $G3$ |
| | 60 | 30.8 | 2700.0 | 11.4 | 9.2 | $G4$ |
| | 15 | 7.7 | 2702.7 | 14.9 | 12.0 | $G5$ |
| | 3 | 1.5 | 2490.2 | 9.5 | 14.3 | $G6$ |
| Shangdong | | | | | | |
| | 11 | 5.7 | 1330.4 | −38.4 | 0.0 | $G0$ |
| | 9 | 4.7 | 2174.2 | 14.0 | 2.4 | $G1$ |
| | 29 | 15.0 | 2607.6 | 12.7 | 4.4 | $G2$ |
| | 48 | 24.9 | 2738.9 | 4.6 | 6.5 | $G3$ |
| | 76 | 39.4 | 2389.6 | −4.6 | 9.1 | $G4$ |
| | 15 | 7.8 | 3195.8 | 11.3 | 12.0 | $G5$ |
| | 5 | 2.6 | 2481.5 | 10.1 | 14.4 | $G6$ |
| Total | | | | | | |
| | 48 | 4.9 | 1358.3 | −31.4 | 0.0 | $G0$ |
| | 43 | 4.4 | 2245.8 | 15.1 | 2.6 | $G1$ |
| | 151 | 15.4 | 1720.7 | −6.8 | 4.3 | $G2$ |
| | 285 | 29.1 | 2119.4 | 4.5 | 6.5 | $G3$ |
| | 337 | 34.5 | 2227.0 | 1.0 | 9.2 | $G4$ |
| | 89 | 9.1 | 2406.9 | 6.5 | 12.0 | $G5$ |
| | 25 | 2.6 | 1884.8 | 5.1 | 14.4 | $G6$ |

Note:
[a] $N$ = number of households; $N$% = number of households as percentage of total; $YF$ = average income per worker; $YFR$ = residual income (after the effects of all other variables removed); $SH$ = household head's education (in years); and $EG$ = education grade. (For the definition of education grades, see figure 2.) Note that all the variables are geometric means for each education group.

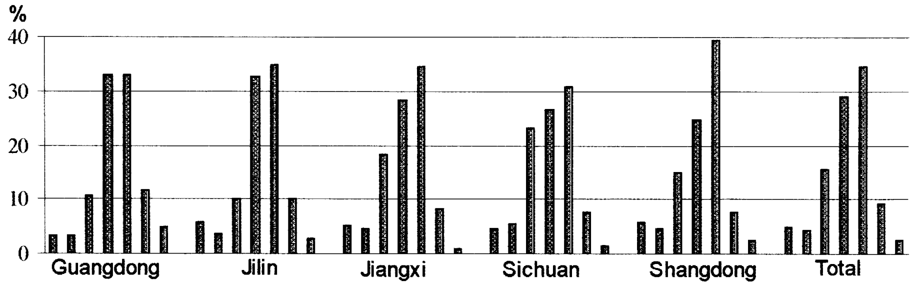


**Figure 1** Distribution of household head's education by household head's education grade

Note: Number of households as a percentage of total. For each country, there are seven bars representing the education grades of household heads in a province, ranging from zero to six as defined in figure 2. The vertical axis gives the percentage of the number of households in a particular education group of a province to the total number of households in the province.
Source: Column 2 of table 5.

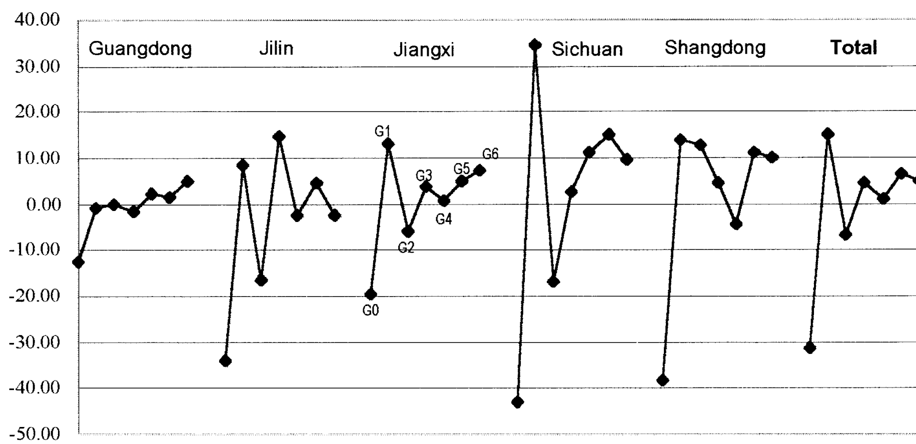


**Figure 2** Percentage deviation of income per worker from mean vs household head's education grade

Notes: [a] $G0$, $G1$, $G2$, $G3$, $G4$, $G5$ and $G6$ refer to the groups of households whose heads received zero, pre-junior primary (1–3), junior primary (4–5), senior primary (6–8), junior middle (9–11), senior middle (12–13), higher education (14–14+) respectively.
[b] Let $Y$ be mean residual household income per worker in a province (e.g. Guangdong) and $Z$ be the mean residual income for an education group (e.g. $G0$), the percentage deviation of household income per worker from mean is given by $100\ (Z/Y - 1)$. Residual income is defined as the income after the effects of all variables other than education are removed. The data for the figure are given in the second column of table 5.
Source: Column 3 of table 5.

have been removed.[12] (Note that these percentages are in terms of the mean income for all households.) Thus, the difference between income per worker of the households in $G0$ and $G1$ amounts to 46 per cent of the mean income per worker for all groups.[13] It can be seen from figure 2 that the percentage difference between the incomes per worker of the first two groups is very large for all the provinces except for Guangdong. Thus, the very large magnitude for the effect of the first three years of education on farm income reflects the abnormally low income of households in group $G0$ (with heads having zero education) and the surprisingly high income of households in group $G1$ (with heads having one to three years of education).

Our regression results clearly imply that the first three years of education of household head yield substantial economic benefits (in terms of higher income per farm worker). However, in view of the high costs of educating farmers in remote and mountainous regions, our regression results should not be accepted without further investigation about possible alternative explanations for the low income for the households with uneducated heads.

Our finding that an additional year of education after the third year of education is associated with a sharp drop in income (table 3, column 4 or 5) is clearly an anomaly, which weakens considerably our regression results as a piece of evidence in favour of productivity gain of farmer education generally. From figure 2, once the point for group $G0$ is removed for each province, an increase in the number of years of schooling does not appear to provide a clear increasing pattern. For Jilin, Jiangxi and Sichuan, raising the household head's education level from three to five years appears to reduce sharply the residual income per worker. The graph for each province in figure 2 suggests that even if education has a favourable effect on income,

[12] To examine the partial effect of $SH$ on $YF$, the conventional procedure would be to plot the mean for each education grade of the residual of $YF$ (from a regression of $YF$ on a set of variables) against the mean for each education grade of the residual of $SH$ (from a regression of $SH$ on the same set of variables). This would produce almost the same graphs as plotting column 2 against column 4 (under $YFR$ and $SH$ headings of table 5 respectively). Removing the effects of all inputs on $SH$ leaves its value largely unchanged because $SH$ is virtually independent of all the variables included. The $R^2$ of the regression of $SH$ on the input variables and the five dummy variables (for the provinces' intercepts) is only 4 per cent, with all the slopes numerically very small and only that of land area per worker is statistically significant. In figure 2, the intervals between every pair of points are the same, because the residuals of $YF$ are plotted against grades (with integer values). Plotting column 2 against column 4 of table 5 would result in irregular intervals between each pair of points but would not alter the height of each point and hence would not produce any new insight.

[13] Let $A$ and $B$ be the average income per worker for households in groups $G0$ and $G1$ respectively and let $M$ be the average income for all households, then the percentage difference between the income per worker of households in $G0$ and $G1$ is given by: $100(A - B)/M$.

this effect is rather weak and is likely to be swamped by the variables which are omitted from the regression.

It can argued that the regression result (that the coefficient of chemical fertilisers is lower for poor households in *G*0 than for the relatively richer households in other education groups) is consistent with both of the following two hypotheses: (i) households in group *G*0 are poor because they have poor land (to the extent that the following proposition A is true: poor return to fertilisers can be regarded as a proxy for poor land); and (ii) households in group *G*0 are poor because their heads have zero education (to the extent that the following proposition B is true: poor return to fertilisers can be attributed to illiteracy). In fact, we have no direct statistical evidence on either propositions A or B. Which proposition is more believable has to be based on some other considerations.

To support the first hypothesis, we can cite the finding from Wu *et al.* (1996) that one of the most significant causes of poverty for households in the regions officially classified as poor (both statistically and numerically) is the unfavourable topography of their land – in the sense that it tends to be in the mountainous areas (Wu *et al.* 1996, table 6, p.13).[14] Hence, it is possible that poverty is the cause of the illiteracy rather than vice versa or that both are the joint effect of land being in the remote and/or mountainous regions, land which is more difficult and costly to irrigate. Although poor land quality (because of unfavourable topography) could be an important cause of poverty for poor households even in regions that are not classified as poor, we cannot directly test the counter-proposition that poor land causes poverty (which in turn causes illiteracy) without information on the quality (or location or topography) of the lands owned by households in our sample. However, it is still important to highlight the possible reasons for the reverse causality between literacy and poverty.

Even though we do not have direct observations on the location of land for our sample of farmers, low return to fertilisers may be evidence of (or

[14] Using a survey sample of 500 households living in four out of six regions in China officially defined as poor regions (the two regions excluded being Mongolia and Tibet), Wu, Richardson and Travers obtain the following equation:

$$\text{Income} = 951^* - 74^*\,\text{Lab} + 0.4\,\text{Land} + .08^*\,\text{Cap} - 126^*\,\text{Topography} - 63\,\text{Illiteracy} + 233^*\text{Deprat} + \sum_i^5 \alpha_i$$

where Illiteracy is the proportion of illiterate adults in the household; Lab, Cap and Deprat are labour, capital and dependency ratio respectively; and the expression $\sum_i^5 \alpha_i$ is the sum of five provinces' intercepts; $^*$ denotes that the coefficient is significant at 1 per cent. All the coefficients are significant except for Land and Illiteracy.

proxy for) mountainous land because mountainous land is difficult or costly to irrigate and an abundance of irrigated water is needed for the effect of fertilisers on the yield of new seeds to reach its full potential.[15] It is well known that a lack of irrigated water would be expected to reduce the returns to chemical fertilisers (when applied to modern seeds) and available evidence from the studies of irrigation in developing countries suggests that irrigated water tends to be more abundant in flat land than in mountainous land because it is cheaper to irrigate flat land. For example, the stream of benefits attributable to the irrigation project in Bangladesh was found to be sensitive to the average area covered by a low-lift pump or tubewell and this area is smaller for land on the hills than on the planes (Nguyen and Alamgir 1976, p. 101). We can therefore test the hypothesis that the coefficient of fertilisers is smaller for the poorest group ($G0$) than for all the other groups.

To carry out this test, using the general-to-specific approach, we have first to make the more general assumption that the coefficients of all inputs are different for the households in group $G0$ than for the rest of the households. Note that a lower coefficient for any other input would not be considered as evidence in favour of the hypothesis that mountainous land causes low income. In other words, we start with a general model with the inclusion of a set of dummy variables to allow for 'shifts' in the output elasticities for all the variables in equation 5 (i.e. $L_{i0} = D_0 L_i$, where $I = ST, CF, LD, CT, T$ and $K$; and $D_0$ is a dummy variable defined as follows: $D_0 = 1$ for households in group $G0$ and $D_0 = 0$, otherwise). After running a series of regressions based on equation 5, we found that only the coefficients of fertilisers and labour days for group $G0$ (with uneducated household heads) are significantly different from the coefficients of these two variables for the other groups (with educated household heads) and that the output elasticity of chemical fertilisers ($CF$) is found to be (numerically and statistically) significantly lower, thus confirming our prior expectation. Until we have direct evidence which indicates that households in group $G0$ do not have land concentrated in mountainous areas, we believe that our regression results do provide some useful evidence in support of the counter-proposition that poor land (rather than illiteracy) causes poverty. In such a case, the observed relationship between poverty and illiteracy can plausibly be explained in

[15] Suppose $x$ causes $y$. It is not uncommon for scientists to use $y$ as a proxy for $x$, if $y$ is unobservable. For example, the presence of a black hole is observed via its effect on its nearby stars. If risk-averse farmers are widely known to specialise in producing only traditional products such as rice rather than cash products such as fruits or animal products, then the ratio of income from traditional products to total farm income can be used as a proxy for risk-aversion, which is not directly observable.

terms of poor farmers being unable to afford the high opportunity and transport costs of sending their children to schools (even if tuition is free).

Note that the coefficient of labour days ($LD$) is significantly higher for group $G0$ (with illiterate household heads) than for the rest of the other groups, probably suggesting that households in group G0 adopt the more traditional labour-intensive farm techniques than households in other groups. Comparing columns 2 and 4 of table 4, it can be seen that the education of household head ceases to have any direct effect, once one (i) allows for households in group $G0$ having different output elasticities with respect to fertilisers and labour time; and (ii) removes the distinction between those in group $G1$ (households whose heads had up to three years of education) and the rest of the sample (i.e. omitting $SH_1$). In other words, our results suggest that hypothesis (ii) holds only if proposition B is true.

To support proposition B, we can argue that the modern technology (which involves the use of chemical fertilisers and new seeds) is complicated to understand and implement so that illiterate household heads are more inclined to make mistakes than their literate counterparts. Against this, we can point out that the Chinese farmers have easy access to written information (from extension services) on what inputs (including fertilisers) to use and how to use them. If they cannot read, their children most probably can. If they are too proud to ask their children, they can ask their neighbours (whom they meet quite frequently) to read for them. In any case, they are rather reliant on directives and so have little freedom to make either mistakes (because of poor education) or improvements (as a result of good education).

## 6. Conclusion

Following a general-to-specific approach to model selection and using the household data relating to a large sample of about 1 000 farms in five Chinese provinces, this article arrives at regression results which suggest that (i) the returns are statistically significant for the education of the household heads but not for the education of farm workers generally; and (ii) the returns are considerably higher for the first three years of education of the household head than for subsequent years. This appears to confirm our household head literacy hypothesis that the household heads, who are decision-makers within each farm, are handicapped in getting the best of the improved seeds and modern agricultural practices if they cannot read or write. It can be argued that this handicap is not overcome even if there are younger members of the households who can read or write, because of the reluctance of the household heads to rely on other members.

However, the above results largely reflect the very low residual income of the households with illiterate heads. It can be argued quite plausibly that literacy *per se* may not be important in the making of farm decisions even in a modernising environment. Technology used in farming is usually simple enough to be passed on orally and the necessary skills and know-how can be transferred from one household head to another by word of mouth (since they do often meet as neighbours) or by the extension workers to each household head by practical demonstration in the fields. In China, local authorities play a particularly important role in providing agricultural extension services. It is not uncommon for local officers to direct farmers to use new seeds and/or new techniques in agriculture and arrange for experts from elsewhere to come and demonstrate new ways of doing things to farmers. This means that even farmers with little or no education can follow the basic instructions satisfactorily in making agricultural decisions such as those on crop mix and input mix. It is not surprising therefore that the weak effect of literacy on income is confirmed by Wu *et al.* (1996, pp. 14–15), on the basis of a sample of 500 rural households officially classified as poor (see the regression equation in footnote 14).

Poor households are more likely to have illiterate household heads because they cannot afford education, which can be expensive for those living in remote farm areas. The households may be poor because their lands are largely in mountainous and/or remote areas and their workers have poor health and less capital to work with, etc. In other words, low income and illiteracy may be positively correlated because they are the joint effects of the same causes – causes which happen to be omitted from the model because they are not observable or because there are no data on them. This study highlights the need for finding out, in some future surveys of households, whether and to what extent (i) illiterate household heads could not attend school because, being poor, their parents could not afford the opportunity cost of sending them to schools when they were young; and (ii) they have poor quality land (e.g. mountainous land with poor soil quality), which helps to explain their low income. By using low response to fertilisers as an indicator of mountainous land, we find some indirect evidence in support of the view that it is poor land which causes households in group $G0$ both to be poor and their heads to be illiterate. Strictly speaking, it is past poverty that was responsible for household heads being illiterate. It can be argued that the same poor land which caused the poverty of the parents of the household head is still the reason for the poverty of the family.

Although our regression analysis using the data from China suggests a considerably larger numerical (and statistically significant) economic effect of education (particularly for the household head) than all the empirical studies included in the two surveys by Lockheed *et al.* and Phillips, we have raised some serious doubts about accepting this conclusion. Our study highlights the need for a re-assessment of the regression results of the previous studies to determine whether (i) the significant economic effect of education is achieved by misleadingly including in the sample a group of outliers (e.g. the group of poor households with illiterate heads in this study); (ii) the high significance level is achieved largely because of regions with heterogenous physical and environmental characteristics are pooled to increase the sample size; and (iii) the low education level is caused by poverty rather than the other way round.

For the case of China, we have raised a number of objections against accepting the very strong regression results in favour of the view that the literacy of the household heads has a substantial impact on household income. This article stresses the need to scrutinise the regression results and significance tests in support of a theory to ensure that the results are not spurious and/or seriously affected by multicollinearity, omitted variable bias or outliers. In so far as some of our objections apply to the regression results in other studies concerning other countries, there is a need for a re-assessment of these results. However, one possible reason for education to play a less important role in terms of productivity gains in some regions of China which may not be true elsewhere is that farmers there have been heavily reliant on directives as to what to produce and how to produce it. Education becomes less important in so far as farmers rely on someone else's judgments rather than on their own. Unless there is a dramatic change in the Chinese government policy, even the next generation of farmers may not be sufficiently free from official directive to make their own decisions so that farmer education may remain ineffective for agricultural productivity gains in the foreseeable future.

Even though the rural environment in China has many characteristics of a modern environment, the lack of empirical support for the productivity gain from farmer education is not really surprising, in view of the fact that Chinese farmers in general have continued to focus most of their efforts on producing a few safe traditional crops such as rice, wheat and maize and are generally unable to accept the higher risks associated with fruits, vegetables, animal and aquatic products and other high-value cash crops. It appears *a priori* that education would become effective only in a really modern environment where farm activities (concerning crop mix, input mix, marketing, finance and insurance) become sufficiently complex. A country's agricultural environment can be classified as modern according to the criteria given in the introduction and yet may offer little advantage to education, unless the farmers also have ready access to the markets for capital (credit), farm input and output to be able to accept the risks associated with the adoption of the innovative planting methods and high-value non-traditional crops and animal products. The simple policy implication of this study is that, when allocating more funds to schools to raise farm efficiency, the Chinese government should speed up the process of modernising the Chinese rural environment. The government can create the modern agricultural environment by deregulating China's markets for major farm input and output, such as the markets for chemical fertilisers, grain and cotton, by abolishing production plans in regional China, and by providing Chinese farmers with easier access to markets for credit, modern inputs and output. All these measures tend to facilitate the production of new farm products and adoption of unfamiliar but efficient techniques by Chinese farmers, thus improving farm productivity.

Of course, improving rural education would yield numerous important social, economic and human welfare benefits other than productivity gains in agriculture. To ensure that the importance of farmer education is duly appreciated, this final paragraph is devoted to a brief discussion of a number of other important benefits of education. First, educated farmers and rural youth (e.g. middle school graduates) are in a better position to find off-farm jobs in both urban and rural areas of China. The continuous migration of educated labour from relatively poor rural areas of China has been found to be one of the main contributors to the high rate of growth in coastal China. Next, educated rural couples usually have less children and thereby help to keep down the birth rate in China. Widespread farmer education also facilitates disease control and immunisation programmes for children (e.g. educated people have better knowledge of how to avoid infectious diseases such as AIDS). Third, the social infrastructure and order can be maintained more effectively if fewer of the rural people are illiterate (since illiterate people cannot read signs for the protection of historical and culture heritage, transport, communication and irrigation equipment). It is more difficult to make uneducated people understand the full extent of the damage their actions can cause (e.g. such as stealing communication and irrigation equipment, shooting endangered animals for small gain). Illiterate people generally have more difficulties obeying laws which they cannot quite understand (this has caused serious problems for themselves as well as for other people; in China, for example, the sales of fake medicine and seeds have cost lives and whole agricultural crops in some areas). It is also found that less educated people tend to settle their disputes by fighting rather than peacefully via the law courts. Poorly educated people are much more inclined to commit offences such as kidnapping children and wife beating. The social and economic effects of education are cumulative, since less educated people tend to have less educated children. Finally, democracy (if it eventually comes to China) can hardly be expected to work properly if the voters are not sufficiently educated to understand what they are voting for.

## References

Crofton, I. 1990, *The Guiness Encyclopedia*; Guiness: London, pp. 58–9.

Gilbert, C.L. 1986, 'Professor Hendry's econometric methodology,' *Oxford Bulletin of Economics and Statistics*, vol. 48, no. 3, pp. 283–307.

Griliches, Z. 1977, 'Estimating the returns to schooling: some econometric problems', *Econometrica*, vol. 43, no. 1, pp. 1–22.

Gujarati D.N. 1995, *Basic Econometrics*, 3rd edn, McGraw-Hill: New York.

Hendry, D. and Richard, J.F. 1983, 'The econometric analysis of economic time series', *International Statistical Review*, vol. 51, no. 2, pp. 111–63.

Kalbfleisch, J.G. and Sprott, D.A. 1976, 'On tests of significance', in Harper, W.L. and Hooker, C.A. (eds), *Foundations of Probability Theory, Statistical Inference, and Statistical Theory of Science,* vol. 2, D. Reidel: Boston, pp. 259–72.

Kennedy, P. 1985, *A Guide to Econometrics*, 2nd edn, The MIT Press: Cambridge, Mass.

Leamer, E.E. 1978, *Specification Searches*, Wiley: New York.

Lindley, D.V. 1957, 'A statistical paradox', *Biometrika*, vol. 4, pp. 187–192.

Lockheed, M.E., Jamison, D.T. and Lau, L.J. 1980, 'Farmer education and farm efficiency: a survey', *Economic Development and Cultural Change*, vol. 29, no. 1, pp. 37–76.

Maddala, G.S. 1992, *Introduction to Econometrics,* 2nd edn, Macmillan: New York.

Mellor, J. 1976, *The New Economics of Growth*, Cornell University Press: New York.

Mitchell, J. and Stein, J. (eds) 1990, *The Random House Encyclopedia,* 3rd edn, Random House: New York.

Nguyen, D.T. and Alamgir, M. 1976, 'A social cost-benefit analysis of irrigation in Bangladesh,' *Oxford Bulletin of Economics and Statistics*, vol. 38, no. 2, pp. 99–110.

Phillips, J.M. 1994, 'Farmer education and farmer efficiency: a meta-analysis', *Economic Development and Cultural Change,* vol. 43, no. 1, pp.149–65.

Schultz, T.W. 1975, 'The value of ability to deal with disequilibria', *Journal of Economic Literature*, vol. 13, no. 3, A92–A96.

Wu, H. 1995, 'A note on the CERU-MoA grain farm household survey in China', in Findlay, C. (ed.), *Collection of Papers Applying the CERU/MoA Survey Data*, Chinese Economic Research Unit, University of Adelaide, S.A., Australia.

Wu, G., Richardson, S. and Travers, P. 1996, *Rural Poverty and its Causes in China*, discussion paper, Department of Economics, University of Adelaide, Adelaide.