Modeling Impacts of Soil Conservation on Productivity and Yield Variability: Evidence from a Heteroskedastic Switching Regression

Gerald E. Shively
Department of Agricultural Economics
Purdue University, West Lafayette, IN 47907-1145
(765) 494-4218 (phone); (765) 494-9176 (fax)
shively@agecon.purdue.edu

Selected paper at annual meeting of the American Agricultural Economics Association
2-5 August 1998, Salt Lake City, Utah.

ABSTRACT. This paper measures the impacts of soil conservation on agricultural productivity and yield variability. A two-stage regression model is used that incorporates both endogenous switching and conditional heteroskedasticity. The analysis corrects for bias introduced when conservation-induced changes in productivity are accompanied by self-selection in the technology adoption process. The paper outlines a strategy for jointly measuring the impacts of soil conservation on yields and yield variability, and demonstrates methods for obtaining consistent and efficient estimates of these impacts using data from low-income farms in the Philippines.

Keywords: Philippines, soil conservation, yield variability, technology adoption, self selection, heteroskedasticity

JEL classification: C35, O33, Q15, Q24

Introduction

Researchers seeking to measure the on-site economic benefits of soil conservation, for example using field-level data on agricultural yields, typically encounter two methodological problems. One, productivity effects of any particular soil conservation measure are potentially correlated with unobserved characteristics of farmers who adopt soil conservation, or with unobserved characteristics of their farms (Shively 1997). As a result, estimates of productivity differences obtained from comparisons of technologies alone are likely to be biased. Two, soil conservation measures have the potential to influence yield variability. From an econometric perspective, this means productivity estimates obtained under the assumption of homoskedastic errors will be inefficient. Furthermore, ignoring the specific form of heteroskedasticity hides the possible impact of soil conservation on yield variability.

Given that land degradation is an important economic and environmental policy problem in many middle- and low-income countries (Blaikie 1985; World Bank 1992), understanding the actual impact of soil conservation measures on resource-poor farms seems critical. Numerous studies suggest that given sufficient time soil conservation measures can reduce rates of soil erosion, increase crop yields, and provide a favorable return on a farmer's investment (e.g. Lutz, Pagiola, and Reiche 1994; Shively 1998). This paper presents a framework for accurately measuring the impact of soil conservation on yield and yield variability, accounting for the influence of latent characteristics of adopters.

Model

Consider a population of farmers, each of whom voluntarily chooses whether to adopt soil conservation. Let the binary variable A_i represent the adoption decision for farmer i, with $A_i = 1$ denoting a farmer who adopts and $A_i = 0$ denoting a farmer who does not. Formally, this implies

a *self-selection* mechanism:

$$\begin{aligned} A_i^* &= \gamma' \mathbf{w}_i + \varepsilon_i, \quad \varepsilon \sim N(0,1) \\ A_i &= 1 \quad if \quad \gamma' \mathbf{w}_i \geq \varepsilon_i \\ A_i &= 0 \quad if \quad \gamma' \mathbf{w}_i < \varepsilon_i \end{aligned} \tag{1}$$

Vector **w** contains variables associated with the self-selection process and vector γ contains coefficients to be identified. By assumption, Prob[A = 1] = $\mathbf{\Phi}(\gamma'\mathbf{w})$ and Prob[A = 0] = 1 - $\mathbf{\Phi}(\gamma'\mathbf{w})$, where $\mathbf{\Phi}$ denotes the standard normal distribution function.

To evaluate the impact of the self-selection process on crop yields, consider a model of agricultural production that relates agricultural inputs to yield. The model accounts for the fact that expected yields and expected yield variance may depend on soil conservation adoption either directly, or implicitly. If y_i represents the yield observed on farm i, then the heteroskedastic production function corresponding to adopters and non adopters is:

$$\begin{aligned} y_i &= g_1(\mathbf{x}_{1i}, A_i) + h_1(\mathbf{x}_{1i}, A_i)\varepsilon_{1i} \quad if \quad A_i = 1 \\ y_i &= g_0(\mathbf{x}_{0i}, A_i) + h_0(\mathbf{x}_{0i}, A_i)\varepsilon_{0i} \quad if \quad A_i = 0 \end{aligned} \tag{2}$$

Vectors $\mathbf{x}_1$ and $\mathbf{x}_0$ contain variables believed to influence expected yield and expected yield variability for adopters and non-adopters, respectively.[1] These may include inputs, farm characteristics, and farmer characteristics. The functions $g_1(\mathbf{x}_1)$ and $g_0(\mathbf{x}_0)$ relate input levels and other factors to yields for adopters and non-adopters respectively. The functions $h_1(\mathbf{x}_1)$ and $h_0(\mathbf{x}_0)$

---

[1] A number of functional forms could be used to investigate the relationship between inputs, outputs, and production risk in equation (2). The approach used here follows Just and Pope's (1979) recommendations for a functional form that imposes as little structure on the risk properties of the arguments as possible. In principal, the additive specification in equation (2) permits increasing, decreasing, or constant marginal yield risk. The following additional conditions on equation (2) are assumed to hold:

$$E(\varepsilon) = 0; V(\varepsilon) = \sigma; E(y) = g(x); V(y) = h^2(x)\sigma;$$

$$and \ \frac{\partial E(y)}{\partial x_i} = g(x_i); \frac{\partial^2 E(y)}{\partial x_i} = g_{ii}(x_i); \frac{\partial V(y)}{\partial x_i} = 2hh_i\sigma \ .$$

describe how inputs influence yield variance, where ε_{1i} and ε_{0i} are production shocks. By assumption, ε_i, $\varepsilon_{1i,}$ and ε_{0i} are distributed trivariate normal, mean zero, with covariance matrix:[2]

$$\begin{bmatrix} \sigma_1^2 & \sigma_{10} & \sigma_{1\varepsilon} \\ & \sigma_0^2 & \sigma_{2\varepsilon} \\ & & 1 \end{bmatrix}.$$

Following Maddala and Nelson (1975) and Maddala (1983) equations (1) and (2) are referred to as a *switching regression with endogenous switching*. The system can be estimated using the two-step procedure associated with Heckman (1979). First, equation (1) is estimated using a bivariate probit model. Estimated probability measures for each observation (expressed as a function of the switch point in the sample) are then computed and retained in the form of the inverse Mills ratio (IMR).[3] Second, production data for adopters and non-adopters are used to estimate the mean and variance components of equation (2).[4] The IMR is included as a regressor in both the mean and variance equations. Controlling for the selection process through inclusion of the IMR in equation (2) is necessary for obtaining unbiased estimates of the coefficients in the yield equation. Furthermore, specification of the stochastic component in equation (2) is required to obtain consistent and efficient estimates of the deterministic component. In general, standard errors for equation (2) obtained from the two-step procedure outlined above must be corrected. Methods used to compute consistent standard errors are discussed below.

---

[2] The normality assumption made here is conventional but not of trivial significance. See Manski (1988).

[3] The probit model is used to construct a measure of the inverse Mills ratio for each farm. The ratio is defined as $\frac{\phi(\gamma'\mathbf{w}/\sigma_1)}{1-\Phi(\gamma'\mathbf{w}/\sigma_1)}$, where ϕ and Φ are density and distribution functions for the normal distribution, and other variables are as defined above. The IMR is a monotone decreasing function of the probability an observation is in the selected sample (Heckman 1979).

[4] Preliminary analysis indicated that parameter estimates derived from an unpooled sample (i.e. separate estimation of the equations in (2) did not differ qualitatively from those obtained in a pooled sample.

Data

Data used in this study were collected between November 1994 and March 1995 from a sample of 89 upland corn plots in Barangay Bansalan, in the Philippine province of Davao del Sur. The site is described by Garcia et al. (1995). The predominant soil conservation strategy used at the site was contour hedgerows.[5] Hedgerow and non-hedgerow plots were measured as part of the survey, as were yields. Production data spanned a calendar year. Observations corresponding to the wet and dry seasons are distinguished via a binary indicator.

Table 1 reports average yields on both an observed per hectare basis (including hedgerow area) and an effective per hectare basis (corn area only), by cropping season. The latter figure serves as the dependent variable in the regressions reported below. Effective yields ranged from 0 to 3000 kg per hectare, with an average of 1410 kg per hectare. Average yield on hedgerow plots (1440 kg/ha) exceeded average yield on traditional plots (1270 kg/ha). Average dry season yield (1070 kg/ha) was significantly lower than average wet season yield (1670 kg/ha). Sample means for independent variables used in the yield regressions are reported in Table 2.

Results

Selection Regression

The selection model was estimated using a ML probit model. Results are reported in Table 3. The dependent variable for the probit model was a binary indicator of whether hedgerows were used on the plot. Explanatory variables in the probit regression included total farm size (in has);

---

[5] Contour hedgerows are defined as a spatially zoned agroforestry practice (Kang and Ghuman 1991). They are constructed as permanent vegetative barriers, typically consisting of grasses or nitrogen-fixing legumes, planted across the width of a field in rows and spaced 5-10 meters apart. The barriers restrict soil and water movement, and annual crops are grown in alleys between the hedgerows. They have been widely promoted throughout Asia, Africa, and Latin America as an effective and low-cost method for maintaining annual crop cultivation on steep fields (Lal 1990).

the quantity of labor available in the household (in man days per hectare); an indicator of farm ownership (measured as the proportion of cultivated land that was held by secure title); plot size (in has); soil depth on the plot (in mm); the age of the plot (in months); and an estimate of the opportunity cost of adoption. All explanatory variables in the probit model were significantly different from zero at a 90% confidence level. Results indicate farm size, available labor, and tenure security were all positively correlated with adoption probability. Adoption probability was negatively correlated with plot size, plot age, soil depth, and the cost of adoption.

Yield Regressions

Table 4 contains results from four yield regressions. All regressions used the natural logarithm of effective yield per hectare as a dependent variable, and a log-linear Cobb-Douglas functional form.[6] Models 1 and 2 are OLS regressions in which only the mean function $g(\mathbf{x})$ was estimated. Models 3 and 4 are heteroskedastic regressions in which both mean and variance functions were estimated by maximum likelihood methods under the assumption of Gaussian errors.[7] In models 3 and 4 the squared residuals from the mean regressions served as dependent variables in the

---

[6] A log-linear Cobb-Douglas model was justified on the basis of a specification test, following MacKinnon, White, and Davidson (1983). The significance of the estimate of the coefficient a in the model $g = b'\mathbf{x} + a[lng - ln(b'\mathbf{x})] + e$ was assessed, where g represents yield, $\mathbf{x}$ is a vector of independent variables, b is a coefficient vector, and e is a regression residual. Patterns of coefficient significance were similar in linear and log-linear regressions. On the basis of the specification test the linear model was rejected in favor of the log-linear model at a 95% confidence level.

[7] Breusch and Pagan (1979) and Glesjer (1969) tests were used to test the null hypothesis of homoskedasticity in the yield function against the alternative of heteroskedasticity. Residuals used in the tests were obtained from a regression of the equation $g = b'\mathbf{x} + e$, where g represents yield, $\mathbf{x}$ is a vector of independent variables, b is a coefficient vector, and e is a regression residual. Tests were applied to the data using two subsets of independent variables consisting of labor, fertilizer, a dry-season dummy variable, soil depth, and a hedgerow indicator. The Breusch-Pagan test allowed acceptance of the null hypothesis of homoskedasticity in both instances, but the Glejser test recommended rejecting the null. Greene (1990) argues the Glesjer test is more powerful than the Breusch-Pagan test within the specific context of the chosen regression model. Therefore, the null hypothesis of homoskedasticity was rejected.

variance regressions. With the exception of the IMR, which appears only in models 2 and 4, identical sets of independent variables were used in all regressions.

Model 1 establishes a basic pattern that is repeated across all models. Labor and fertilizer are both associated with increases in agricultural yields at statistically significant levels. In elasticity terms, a one-percent increase in available labor was associated with a 0.43 percent yield increase at the mean, and a one-percent increase in available fertilizer was associated with a 0.10 per cent yield increase. Controlling for input use and other factors, dry-season yields were statistically lower than wet-season yields. Based on results from Model 1 and an average hedgerow land share of 10%, one concludes that the presence of hedgerows was associated with an increase in corn yield of approximately 125 kg/ha. However, as the intensity of hedgerow use increased (as measured by the share of land occupied by hedgerows rises), the impact of hedgerows on yield diminished. This pattern likely reflects crop competition for light and water when hedgerow spacing is narrow.

Results from Model 2 indicate that, with one important exception, the results of Model 1 are invariant in sign, magnitude, and significance to the inclusion of the inverse Mills ratio.[8] The important exception is that by accounting for latent characteristics underlying the selection process, the statistical significance of the soil conservation indicator declines. Although the coefficient for the inverse Mills ratio is not significantly different from zero, its inclusion in the yield equation reduces the explanatory power of the hedgerow variable, suggesting that the measured impact of soil conservation is partly embodied in the characteristics associated with

---

[8] Including the IMR in the 2nd stage OLS is insufficient to provide asymptotically consistent standard errors in 2nd stage because the disturbance term in the 2nd stage problem is heteroskedastic by construction. Greene (1981) outlines a procedure for correcting OLS standard errors which involves weighting the OLS variance-covariance matrix with covariances of the probit model coefficients. The correct estimator used for Model 2 was derived from Greene's recommended guidelines using Jaeger's Shazam code.

hedgerow adoption. In other words, a conjecture that hedgerow effects are independent of the self-selection process is rejected with this model.

Turning to the issue of yield variability, results from the variance regressions of models 3 and 4 indicate, not surprisingly, that labor is a risk reducing input. The model also shows that fertilizer is a risk-reducing input on upland farms. For these farms the nitrogen response of corn appears to be greater and more wide-ranging during the wet season than the dry season. Lower conditional yield variance during the dry season is a natural byproduct of this relationship.

Concerning the impact of soil conservation on yield variability, the binary hedgerow indicator in models 3 and 4 demonstrates that hedgerows reduce yield variance slightly. However, the statistical significance of this result hinges on whether one accounts for the sample selection process. Inclusion of the IMR reduces the magnitude and statistical strength of the hedgerow indicator in the variance equation. Including the IMR in Model 4 reduces the magnitude and statistical significance of the hedgerow indicator in the variance equation, and also reduces the statistical significance of other coefficients in the variance equation.[9]

Discussion

Important findings are summarized as follows. First, the measured impact of hedgerows on yield is highest in the OLS model. The OLS regression predicts an average yield that is 1% higher than the OLS switching model, 4% higher than the heteroskedastic model, and 10% higher than

[9] Correcting the standard errors in Model 4 is more problematic than in its homoskedastic counterpart. In particular, no convenient direct correction is available. As an alternative, a bootstrap procedure based on Efron and Tibshirani (1986) was employed to obtain asymptotically consistent standard errors. The bootstrap procedure relied on 1000 replications of a constructed regression in which dependent variables were computed using random sampling with replacement from the observed residual vector. For more on bootstrap methods, see Efron and Tibshirani (1993).

the heteroskedastic switching regression. In other words, the bias introduced by parsimoniously specified yield regressions is as great as 10%.

Second, relative to predicted yields for plots without hedgerows, the predicted yield differentials associated with hedgerows are greatest in the heteroskedastic regressions. For example, models 3 and 4 indicate relative yield differentials of 10% and 8% for hedgerow plots compared with 6% and 4% for their homoskedastic counterparts. These results suggest that at least part of the yield impact of hedgerows may be unaccounted for when the determinants of yield variability are ignored. Third, one finds smaller and less significant yield differentials when accounting for latent characteristics of adoption (models 2 and 4 vs. models 1 and 3). This supports a conjecture that observed yield differentials reflect the underlying characteristics of adopters or their plots, rather than soil conservation, *per se*.

Conclusions

This paper presented a framework for estimating yield impacts of soil conservation using a heteroskedastic production function with endogenous switching. Data from hillside farms in the Philippines were used to test the hypothesis contour hedgerows influence yields. Findings suggest hedgerows were associated with higher yields and lower yield variability. However, the magnitude and statistical strength of these relationships were found to depend on the choice of regression model. Information on latent characteristics of adopters diminishes the statistical strength of soil conservation parameters in yield regressions. An OLS model was found to overstate yield impacts compared with a heteroskedastic model that accounted for latent characteristics of adopters via a first-stage probit model. These findings have implications for the practical application of soil conservation strategies in low-income settings. Future research should focus attention on further distinguishing factors that help explain yield variance.

References

Blaikie, P. 1985. *The Political Economy of Soil Erosion in Developing Countries*. London: Longman.

Breusch, T. and A. Pagan. 1979. "A Simple Test of Heteroscedasticity and Random Coefficient Variation." *Econometrica* **47**:1287-1294.

Efron, B. and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Efron, B. and Tibshirani, R. 1986. "Bootstrap Measures for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science* **1**(1):54-77.

Garcia, H. N. M., et al. 1995. Soil Conservation in an Upland Farming System in Davao del Sur: A Socio-economic survey. SEARCA-UQ Upland Research Project Report No. 2. Los Banos: Southeast Asian Regional Center for Graduate Study and Research in Agriculture.

Glesjer, H. 1969. "A New Test for Heteroscedasticity." *Journal of the American Statistical Association* **60**:539-547.

Greene, W. H. 1981. "Sample Selection Bias as a Specification Error: Comment." *Econometrica* **49**(3):795-798.

Greene, W. H. 1990. *Econometric Analysis*. New York: MacMillan.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* **47**(1):153-161.

Jaeger, D. A. 1992. Sample Selection-Correction Estimation ("Heckit"): SHAZAM Heckman Correction Program.

Just, R. E. and R. Pope. 1979. "Production Function Estimation and Related Risk Considerations." *American Journal of Agricultural Economics* **61**(1):277-284.

Kang, B. T. and B. S. Ghuman. 1991. "Alley Cropping as a Sustainable System." In W. C. Moldenhaur, N. W. Hudson, T. C. Sheng, and S. W. Lee (eds.). *Development of Conservation Farming on Hillslopes*. Ankeny: Soil and Water Conservation Society.

Lal, R. 1990. *Soil Erosion in the Tropics-Principles and Management*. New York: McGraw-Hill.

Lutz, E., S. Pagiola, and C. Reiche. 1994. "The Costs and Benefits of Soil Conservation: The Farmer's Viewpoint." *The World Bank Research Observer* **9**(2):273-295.

MacKinnon, J. H. White, R. Davidson. 1983. "Tests for Model Specification in the Presence of Alternative Hypotheses: Some Further Results." *Journal of Econometrics* **21**(1):53-70.

Maddala, G. S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maddala, G. S. and F. Nelson. 1975. "Switching Regression Models with Exogenous and Endogenous Switching." *Proceedings of the American Statistical Association* 423-426.

Manski, C. F. 1988. "Identification of Binary Response Models." *Journal of the American Statistical Association*. **83**:729-738.

Shively, G. E. 1997. "Consumption Risk, Farm Characteristics, and Soil Conservation Among Low-income Farmers in the Philippines," *Agricultural Economics* **17**(1997):165-177.

Shively, G. E. 1998. "Impact of Contour Hedgerows on Maize Yields in the Philippines." *Agroforestry Systems* **24**(1):159-168.

World Bank. 1992. *Development and the Environment, World Development Report 1992*. Washington, DC: The World Bank.

Table 1. Sample means for corn yields

	Wet season	*Dry season*	*Traditional plots*	*Hedgerow plots*	*All plots*
Observed	1670	1070	1270	1440	1340
Per hectare	1770	1130	1270	1620	1410
number of observations	50	39	53	36	89

Table 2. Sample means for independent variables used in yield regressions

	Average per hectare				*Average per effective hectare*			
	All plantings	Second planting	Non-hedgerow	Hedgerow plots	All plantings	Second planting	Non-hedgerow	Hedgerow plots
% of plots with hedgerows	0.39	0.40	-	1.0	0.39	0.40	-	1.0
% of plot occupied by hedgerows	0.05	0.06	-	0.12	0.05	0.06	-	0.12
Labor (days/ha)	326	334	267	412	352	366	267	477
Nitrogen fertilizer (kg/ha)	136	141	130	145	146	151	130	170
Age of plot (months)	83	85	90	117	83	85	90	117
Slope (degrees)	26	26	25	27	26	26	25	27
Soil depth (mm)	850	847	838	867	850	847	838	867
n	89	50	53	36	89	50	53	36

Table 3. Results of probit analysis of soil conservation adoption

Independent variable	*Coefficient estimate*	*Standard error*
Constant	1.1722	1.9255
Farm size (hectares)	0.1666	0.0821*
Available household labor per hectare (man days per hectare)	0.0021	0.0011*
Proportion of cultivated area with secure tenure (0,1)	1.2737	0.6700*
Plot size (hectares)	-0.9187	0.5200*
Soil depth of plot (mm)	-0.0028	0.0014*
Period of continuous cropping on plot (months)	-0.0162	0.0084*
Ratio of initial cost of adoption on plot to total household corn availability	-0.4498	0.1663*
Value of log-likelihood function†	-48.71	
Proportion correct predictions	0.65	
number of observations	89	

* Coefficient is statistically different from zero at the 90% confidence level.
† Likelihood ratio test for regression with constant only is -60.1.

Table 4. Yield equation results

Mean Equation: dependent variable is natural logarithm of effective yield per hectare

Independent variables	1: OLS	2: OLS w/Probit†	3: Heteroskedastic	4: Heteroskedastic w/Probit††
Constant	4.6317* (0.5490)	4.6235* (0.5902)	5.3231* (0.3365)	5.1999* (0.4688)
Hedgerows (0/1)	0.3914* (0.2208)	0.3766 (0.4962)	0.4157* (0.1293)	0.4241 (0.3830)
Log of labor (man days per hectare)	0.4281* (0.0977)	0.4307* (0.1191)	0.3439* (0.0537)	0.3466* (0.0955)
Log of fertilizer (kg/ha)	0.1047* (0.0345)	0.1048* (0.0350)	0.0488* (0.0170)	0.0606* (0.0261)
Dry season (0/1)	-0.5213* (0.1724)	-0.5210* (0.1736)	-0.5268* (0.0840)	-0.5135* (0.1308)
Hedgerow share	-3.3195 (2.2540)	-3.3400 (2.3270)	-3.2404* (1.1750)	-3.5037* (1.8768)
Inverse Mills ratio from adoption probit	–	0.0120 (0.3055)	–	-0.1580 (0.2395)
Variance Equation: dependent variable is squared residual from mean equation				
Constant	–	–	1.0857* (0.2572)	1.0141 (0.6321)
Hedgerows (0/1)	–	–	-0.1942* (0.0924)	-0.0056 (0.0057)
Log of labor (man days per hectare)	–	–	-0.1325* (0.0392)	-0.1445 (0.1143)
Log of fertilizer (kg/ha)	–	–	-0.0353* (0.0101)	-0.0365 (0.0467)
Dry season (0/1)	–	–	0.1652* (0.0494)	0.2231 (0.2687)
Hedgerow share	–	–	2.3652* (0.7738)	2.4499 (2.6286)
Inverse Mills ratio from adoption probit	–	–	–	-0.1275 (0.3348)
R^2	0.33	0.33	–	–
Log of likelihood function	–	–	-85.21	-84.89
number of observations	89	89	89	89

* Coefficient is statistically different from zero at the 90% confidence level.
† Standard errors directly corrected for selection-induced heteroskedasticity. See text.
†† Standard errors derived from a bootstrap procedure. See text.