Impact Assessment of Technologies that Mitigate Adverse Circumstances: The Case of Disease Resistant Beans in Honduras

David L. Mather
Ph.D. Candidate
Department of Agricultural Economics
Michigan State University
matherda@msu.edu

James F. Oehmke
Associate Professor
Department of Agricultural Economics
Michigan State University
oehmke@msu.edu

Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Montreal, Canada, July 27-30, 2003

1. Introduction

Conventional *ex post* impact assessment methods have typically focused on productivity enhancement technologies, whose benefits are estimated as the yield difference between traditional and improved technologies as observed in either experimental trials or a cross-section survey of farmers. However, both of these data sources have limitations in constructing the counterfactual situation used to assess the farm-level benefits of productivity maintenance technologies. Various authors have used experimental trial data to estimate yield loss averted through use of a disease-resistant cultivar (Smale *et al* 1998*,* Morris *et al*, 1994), but experimental trials often do not well-approximate farmer conditions. This is especially true for measuring disease resistance technologies, considering that disease frequency and intensity are fixed in experimental trials, yet under farmer conditions they may vary spatially by weather patterns, altitude, and crop management practices. However, the presence of selection bias in farmer survey yield data will tend to underestimate the real benefits of disease-resistant technologies, to such an extent that Johnson and Klass (1999) recommend the use of experimental trials over survey data, even with the limitations of the trial data.

This essay proposes methodological advances that allow for the use of farm survey data to address the questions of yield loss avoidance and returns to research. The point of departure is Otsuka *et al* (1994), who use farm-level survey data to estimate the impact of RV (disease-resistant variety) rice in the Philippines. Otsuka *et al* apply a modified Heckman two-step procedure (adapted from Lee, 1978) to correct for selection bias. Lee's procedure gives unbiased and consistent estimates of parameters of the equations of interest (traditional variety (TV) yield and RV yield), which Otsuka uses to compare the mean rice yields of TVs and RVs.

However, these equations represent the yield response of a given subsample of the population – TV growers in one case, and RV growers in the other. In other words, it compares the expected RV yield (conditional on RV use) with the expected TV yield (conditional on TV use). However, the more appropriate counterfactual scenario for impact analysis is unobservable – the yield that RV users would have obtained in the absence of RVs.

Extending Otsuka *et al*, this paper constructs the counterfactual to RV yields as what yields RV users would have obtained had they continued to grow TVs (i.e. the TV yield conditional on RV use). Constructing the appropriate counterfactual is especially poignant in the case of RVs (and other technologies that mitigate adverse circumstances), since it is anticipated that the TV yield profile differs significantly between RV users and TV users. Specifically, farmers adopt RVs to avoid yield losses from disease. Farmers who choose to remain with the TVs must not experience significant yield losses, otherwise they would adopt the RV. Using farm-level survey data on disease-resistant bean yields in Honduras, a modification of Lee's two-step procedure is employed to obtain selection-corrected bean yield equations for TV and RV users. As demonstrated in labor supply literature (Lee, 1978; Duncan and Leigh, 1980), estimation of these selection-corrected equations enables the prediction of imputed yields for each individual's unobserved varietal choice, conditional on his/her own observed characteristics. That is, the selection-corrected yield equations account for the endogenous farmer decision to adopt or not adopt RVs, and allows for meaningful prediction of TV yields (including yield loss) that those farmers adopting RVs would have gotten had the RVs not been available. Thus, the predicted TV yields of RV users can be statistically compared with predictions of RV yields to estimate the farm-level impact of disease-resistant bean varieties in

Honduras, and the impact of the research that developed these varieties.

2. Model

Our principal interest in this paper is to estimate the impact of RVs by predicting the counterfactual yield of RV users -- the TV yield that RV adopters would have gotten in the absence of RVs -- and comparing this with predictions of RV yields for the same subgroup.[1] In the following model, each farmer's observed yield depends on his varietal choice, represented by equation (2). Yield Y_{Ti} is observed if the farmer grows a TV, while Y_{Ri} is observed if the farmer grows an RV. An implicit assumption of the model is that farmers have a good idea of their disease pressure.

$$I_i = \gamma_0 + \gamma_i Z_i + e_i \qquad (1)$$

We observe:

Y_{Ri} when $I_i = 1$ if farmer grows a RV

Y_{Ti} when $I_i = 0$ if farmer grows a TV, but never both

$$\ln Y_{Ri} = B_{R0} + B_{Ri} X_{Ri} + u_{Ri} \qquad (2)$$

$$\ln Y_{Ti} = B_{T0} + B_{Ti} X_{Ti} + u_{Ti} \qquad (3)$$

where: $\ln Y_{Ri}$ and $\ln Y_{Ti}$ are the logs of yield for individual i, X_i is a vector of personal and

[1] Comparing predicted TV yields with predicted (instead of actual) RV yields ameliorates the influence of measurement error on the actual yields.

farm-level characteristics, and u_{Ri} and u_{Ti} are residuals. This is a switching regression model with endogenous switching, as first developed by Goldfeld and Quandt (1973) and later modified by Lee (1978).

3. The sample selection problem

Consider a population of farmers who can be identified as either adopters or non-adopters of RVs. A random sample survey of this population will result in two subsamples of farmers: one with observations on RV yields, the other with observations on TV yields. Whether or not sample selection creates a problem for prediction of varietal yields from this survey data depends upon whether or not there are differences in both the observable and unobservable characteristics of TV and RV growers.

First, assume that the decision to adopt an RV is random, and that the RV subsample has similar endowments of characteristics as the TV subsample. In this case, there is no reason to suspect selectivity bias will be induced by examining the RV subsample (or vice versa), because the average characteristics (observables and unobservables) of RV users should be similar to the average characteristics of the population (Vella, 1998).

However, once we assume that the decision to adopt an RV is not random, then the RV and TV subsamples potentially have different characteristics. Sample selection bias in this case occurs when some component of the varietal choice decision is relevant to the yield determination process; that is, when some of the determinants of the varietal decision also influence yield. Yet, when the relationship between the varietal decision and the yield determination is purely through the observables, then appropriate variables in the yield equation can control for this. Thus, sample selection bias will not occur simply because of differences in

observable characteristics (*ibid*, 1998).

When one further assumes that unobservable characteristics (such as disease pressure) affecting the varietal decision are correlated with the unobservable characteristics affecting the yield determination process, then a relationship exists between the varietal decision and the yield determination process. If these unobservable characteristics are correlated with the observables, then the failure to include an estimate of the unobservables will lead to incorrect inference regarding the impact of the observables on yields. Thus, under these circumstances, a bias will be induced due to sample selection (*ibid*, 1998).

In our case, we do not observe RV (TV) yields for some farmers due to the outcome of another variable - varietal choice. Because disease pressure – and perhaps other unobservables which influence varietal choice – is not distributed randomly throughout the sample and is correlated with yield, then $E(u_i \mid \text{varietal choice} \ldots 0)$. If some part of the yield distribution is cut off, then estimated residuals from this model are truncated normal, and OLS estimates of yield equations are inconsistent.

To see the effects of sample selection bias more clearly, consider the following data generating process for yield:

$$\text{yield}_i = B_0 + B_1 * \text{rainfall}_i + B_2 * \text{fertilizer}_i + B_3 * RV_i + B_4 * \textit{disease pressure}_i + u_i \qquad (1)$$

where: RV = 1 if farmer i grows a resistant variety, and RV=0 if not; rainfall and fertilizer use are observed variables; and $\textit{disease pressure}_i$ is an omitted variable

In the absence of a farm-level measure of the true intensity, timing or frequency of disease pressure, and assuming that disease pressure is not distributed randomly throughout the sample, then the results of the yield regression (1) will be biased because the error term is actually $u_i + B_4$*$disease\ pressure_i$. Thus, the mean error will not be zero. In addition, we will encounter a further source of bias in that the omitted variable is likely correlated with the RV dummy.

Roy (1951) provided an early discussion of the problem of self-selectivity, discussing the problem of individuals choosing between two professions, hunting and fishing, based on their productivity in each activity. The observed distribution of incomes of hunters and fisherman was determined by these choices. The later discussion of the econometric consequences of self-selection bias began with studies of womens' participation in the labor force and implications for the estimation of wage equations (Gronau, 1974; Lewis, 1974). The basic idea of the two-stage estimation procedure as first proposed by Heckman (1976) (and later extended by Lee (1976)) is to use information from a probit function of labor force participation to construct a regressor for the wage equation which serves to adjust the wage equation error term so that its expected value will be zero.

4. Estimation Procedure

The two-stage estimation procedure for each yield equation proceeds as follows. First, using all *n* observations, we estimate by ML a probit model of Z_i on I_i to obtain estimates of γ_i and γ_0.

$$I_i^* = \gamma_0 + \gamma_i Z'_i + e_i^* \qquad (4)$$

Conditional on RV use, the RV yield equation is:

$$\ln Y_{Ri} = B_{R0} + B_{Ri}X_{Ri} + \sigma_{Rei}\lambda_{Ri} + v_{Ri} \qquad \text{for Ii} = 1 \qquad (5)$$

where $E(v_{Ri} \mid Ii = 1) = 0$. The term $\lambda_{Ri} = [-f(\psi_i) / F(\psi_i)]$ is known as the inverse Mills' ratio (IMR), where *f* is the density function of a standard normal random variable, and *F* is its cumulative distribution, and where $\psi_i = \gamma_0 + \gamma_1 Z'_{i.}$. The IMR term models the truncation effect on yields associated with sample selectivity, and thus enables $E(v_{Ri} \mid Ii = 1) = 0$. The coefficient on the IMR term σ_{Re} is the covariance between e_i , the error term from the RV selection equation, and u_{Ri}, the error term of the RV yield equation. The actual test for selection bias in the RV yield equation is simply a t-test of whether H_o: $\sigma_{Rei} = 0$ or not. Rejection of the null indicates the presence of selection bias.

Conditional on TV use, the TV yield equation is:

$$\ln Y_{Ti} = B_{T0} + B_{Ti}X_{Ti} + \sigma_{Tei}\lambda_{Ti} + v_{Ti} \qquad \text{for Ii} = 0 \qquad (6)$$

where $E(v_{Ti} \mid Ii = 0) = 0$. The term $\lambda_{Ti} = [\, f(\psi_i) / (1 - F(\psi_i))]$.

The error term e_i in the selection equation is assumed to be distributed standard normal, and also independent of Z_i (and therefore X_i). In addition, X_i must be a strict subset of Z_i, and we must have some elements of Z_i that are not also in X_i. This means that we need a variable that

affects selection but does not have a partial effect on yields (Wooldridge, 2000)[2]. Yield equations in switching regressions are commonly estimated either jointly (pooled) (Otsuka, 1994) or separately (Lee, 1978; Duncan and Leigh, 1980). This paper estimates the two yield equations pooled together to gain efficiency in parameter estimation. An RV dummy serves as an intercept shifter for RV users, and is interacted with some of the regressors. Thus, the expected yield for an RV user is represented by (7) while that of a TV user is represented by (8).

$$\ln Y_{Ri} \mid_{Ri} = B_{T0} + B_{R0}*RV + B_{Ti}X_{Ri} + B_{Ri}X_{Ri}*RV + \sigma_{Rei}\lambda_{Ri} + v_{Ri} \quad \text{for RV=1; Ii = 1} \quad (7)$$

$$\ln Y_{Ti} \mid_{Ti} = B_{T0} + B_{Ti}X_{Ti} + \sigma_{Tei}\lambda_{Ti} + v_{Ti} \quad \text{for RV=0; Ii = 0} \quad (8)$$

In the case of disease-resistant beans in Honduras, TV yields are expected to be truncated from below due to selection bias. That is, low TV yields caused by disease pressure are unobserved if farmers in high-disease areas switch to RVs. Thus, assuming that TV yields depend upon the level of disease pressure, then unobservable characteristics (disease pressure) will likely influence both the varietal decision (with no disease pressure, farmers prefer the TV due to its better price) and the yield determination of TVs. However, it is not clear whether or not the presence (or not) of disease pressure would affect RV yields. That is, RV yields under farmer conditions are expected to be about the same with or without disease pressure, *ceteris paribus*. On the other hand, given that RV yields are not often observed in non-disease areas

[2] If z = x, then λi can be highly correlated with the elements of xi, and such multicollinearity may lead to very high standard errors for the Bi. Thus, without this exclusion restriction, it is extremely difficult to distinguish sample selection from a misspecified functional form for yields (Wooldridge, 2000).

(higher elevations) due to their price discounts relative to TVs, and that RVs are likely better-suited for the conditions of valley farmers, there may well be a part of the RV yield distribution that is not observed.

Because the error term of the pooled regression has a heteroskedastic structure, we apply Huber/White's procedure to obtain robust standard errors. The final step is to calculate predictions of RV users' RV yields using the corrected RV equation (9). The counterfactual prediction is RV users' TV yields, which are calculated using the characteristics of RV users evaluated with the coefficients from the TV equation (10).

$$\ln Y_{Ri} \mid_{Ri} = B_{T0} + B_{R0}*RV + B_{Ti}X_{Ri} + B_{Ri}X_{Ri}*RV + \sigma_{Rei}\lambda_{Ri} \qquad \text{for RV=1; Ii = 1 (9)}$$

$$\ln Y_{Ti} \mid_{Ri} = B_{T0} + \quad + B_{Ti}X_{Ri} + \quad + \sigma_{Rei}\lambda_{Ri} \qquad \text{for RV=0; Ii = 1 (10)}$$

5. Background and Data

Honduras' two bean breeding programs, which are implemented by Zamorano (Programa de Investigaciones en Frijol, Escuela Agricola Panamericana) and DICTA (Direccion de Ciencia y Tecnologia Agricola, the Honduran National Bean Program), work in collaboration with three organizations: the USAID-funded Bean/Cowpea Collaborative Research Support Project (CRSP); the Programa Cooperativo Regional de Frijol para Centroamerica, Mexico, y El Caribe (PROFRIJOL), funded by COSUDE, the Swiss Agency for Development and Cooperation; and CIAT (Centro Internacional de Agricultura Tropical). Bean Golden Yellow Mosaic Virus (BGYMV) is the principal bean disease in Honduras, and one of the main production constraints in Honduran valleys (Martel, 1995). The virus is transmitted by the whitefly species Bemisia

tabaci, which is normally found below 1,000 m in all growing regions of Honduras and more frequently in the drier postrera season (September/ October to December/January). Zamorano and DICTA have developed several varieties with resistance to BGYMV (RVs), the first of which was released in 1990. However, the economic gains from yield loss averted through RV use are at least partially offset by price discounts in the market from 7 to 15 percent (Martel, 1995; Mather *et al*, 2003).

This analysis in this paper is based upon a farm-level survey of bean farmers (N=210) implemented by the CRSP and PROFRIJOL in January-February 2001, in collaboration with Zamorano and DICTA. The survey targeted the Mideast (El Paraiso and Francisco Morazan departments) and Northeast (Olancho) regions, which together account for about one-half of annual bean production in Honduras. Bean producers in Honduras are predominantly small farmers, the majority of whom use fertilizer and insecticide, market at least some of their bean production, and depend upon beans for a major portion of their household income (Martel, 1995; Mather *et al*, 2003). Honduran farmers typically plant two crops during the year. In the primera (May/June to July/August), the rainy season, maize is the principal crop, and beans are either intercropped with maize or monocropped[3]. The postrera (September/October to December/January) is a drier season, during which beans are the major crop and are almost exclusively monocropped. Because the whitefly population is often highest under drier conditions, BGYMV is most frequent and severe in the postrera.

Most of the sample farmers (70 percent) live in villages which have had BGYMV

[3] In the following analysis, intercropped area is converted to "effective monocrop bean area" given information from each respondent on his/her intercropping system.

pressure in the past (Table 2), and also (58 percent) live in villages which have received extension at the village level including promotion, demonstration, and/or access to RVs. RV adoption is widespread, and mean RV yields tend to be higher than those of TVs (Table 1). Yet, because of the high variance of both RV and TV yields, yields by variety are not statistically different.

The basis of impact analysis of varietal development is the estimation of the incremental change in benefits received by the participants -- those who have adopted the variety. In our case, we want to estimate the current yields of RV users, as well as what yields they would have received in the absence of RVs. However, as mentioned above, constructing the RV counterfactual as the current TV yields of TV users ignores the potential for selection bias to underestimate the benefits to RVs. Therefore, assuming that selection bias is present, the appropriate counterfactual to current RV yields is unobservable: the yields these RV growers would obtain if they were to use TVs instead. The fact that the RV survey yields and coefficients of variation are quite similar to those of TVs – given that significant price discounts exist, most farmers market beans, and adoption rates are fairly high – suggests that selection bias is indeed present, and that using current TV yields of TV users would underestimate the benefits of RVs to RV users.

Table 1. Adoption of RVs and Bean Yields by Season, 1999-2000, Honduras

Variety	Primera		Postrera	
	1999	**2000**	**1999**	**2000**
Varietal Use: RVs (% farmers)	45	45	41	42 [a]
TVs	64	62	67	76
Both an RV and a TV	9	7	8	18
Sample size (farmers)	N=164	N=170	N=202	N=202
Yield: RVs (mean kg/ha)	857	769	667	446
TVs (mean kg/ha)	678	632	615	459
Disaster (%)	4	6	6	14
Rainfall; first month of season (avg mm)	126	171	314	206
Sample size (parcels)[b]	N=188	N=203	N=242	N=268

Source: CRSP/PROFRIJOL Farmer Survey, 2001 (N=210)
[a] As some farmers planted more than one variety, the total is greater than 100%.
[b] Parcels aggregated at the farmer and variety level

Table 2. Characteristics of sample farmers in Honduras, 2000

Characteristics	Primera	Postrera
Bean area (mean ha)	1.30	1.58
Altitude (m)	945	913
% Farmers: in Area of Bean Golden Mosaic Virus [a] (%)	70	70
Used formula fertilizer (%)	55	48
Used Urea (%)	49	45
Used both formula and urea (%)	29	25
Intercropped (%)	14	15
Ever planted a resistant variety (%)	67	68
% farmers living in village with extension (by type): [b]		
DICTA artisan seed program (%)	22	19
Dept. of Natural Resources (%)	35	32
Zamorano (%)	16	14
Catholic Church (%)	13	17
No extension at village level (%)	42	45
Sample size	N=170	N=202

Source: CRSP/PROFRIJOL Farmer Survey, 2001 (N=210)
[a] defined by Dr. Juan Carlos Rosas, Director, Programa de Investigaciones en Frijol, Zamorano [b] extension which includes promotion, demonstration, and/or extension of improved varieties

6. Estimation results of varietal choice function

We estimated the probit RV adoption function using data from four consecutive seasons: the primera and postera seasons of 1999 and 2000. Most of the elements of the adoption function are factors expected to influence bean yield such as fertilizer use, rainfall, cropping system, season, etc., as well as village dummies which help to capture non-observable agronomic factors common at the village level. Farmer socioeconomic factors such as age, education, and

farmsize are not included here due to their insignificance.

Formula fertilizer use is defined as zero if the farmer used no formula fertilizer in a given season, and is defined as one if the farmer used formula in that season[4]. Variables for urea fertilizer use and combined formula and urea use are similarly constructed. Rainfall represents the total village rainfall (mm) in the first month of the given season. While rainfall the first month of a bean season is undoubtedly an important factor in yields, the explanatory power of this rainfall data is questionable for several reasons. First, only nine rainstations across the three departments had complete data during the time periods of our four seasons of production data. Thus, each of the 30 villages were assigned the nearest rainstation to represent village rainfall. Second, the timing and distribution of rain within a month is not captured by a monthly total. Intercrop refers to any cropping system which is not monocrop (but does not include the relay system common in Olancho in the postrera, which is essentially monocrop). Disaster is a dummy to capture the influence of yields less than 80 kg/ha.

The remaining factors in the adoption function serve as exclusion restrictions – variables that affect selection but do not have a significant partial effect on yields[5]. These include a dummy variable for villages in which BGYMV has been a problem in the past[6], village altitude

---

[4] We only have fertilizer use data for primera and postrera of 2000. Since this variable measures only use and not fertilizer level, and since many farmers regularly use fertilizer, we assume that fertilizer use in 1999 to be the same as that observed in 2000. This assumption does not affect the test for selection bias, nor the significance or magnitude of key variables in the corrected yield equation such as RV and fertilizer use.

[5] These exclusion restriction variables were included in an OLS yield regression and determined to be insignificant at the 0.10 level.

[6] BGYMV area classification designated by Dr. Juan Carlos Rosas, Director, Programa de Investigaciones en Frijol, Zamorano.

(lower altitudes experience higher BGYMV pressure), village altitude squared, and a dummy variable for farmers who have planted an RV at some time in the past (PRV).

While a measure of a farmer's information about and access to RVs may be thought to be well captured by village-level extension variables, these variables are not ideal exclusion restriction variables since extension presence in a village -- which in this case includes more than just information about and access to RVs – would be expected to influence (increase) yields. BGYMV is only a rough indicator of disease pressure, because what is more significant in the adoption decision is the intensity of disease attack and its frequency. Altitude is also a rough indicator of disease pressure, as lower altitudes tend to have higher BGYMV pressure. PRV is an ideal exclusion restriction variable because it is intimately related to adoption, yet does not predict adoption perfectly due to disadoption (67 percent of the sample has planted an RV at some point, while 45 percent continue to grow an RV) and is not correlated with yields.

The estimation results show that the coefficient on the BGYMV dummy is positive, significant and of relatively large magnitude, which confirms the expectation that the principal benefit of RVs is expected to be the yield loss averted due to their disease resistance (Table 3), and that farmers outside the BGYMV area would not likely adopt the older RVs which have significant price discounts. Also as expected, the coefficient on the PRV dummy is positive, significant and of sizeable magnitude. Altitude is also significant, yet it's sign is contrary to expectation as the probability of adoption would be expected to fall as altitude increased.[7] None of the exclusion restriction variables are significant when tested in the yield regression.

[7] While village altitude is negatively correlated (pairwise) with RV adoption, it's sign becomes positive when village dummies are included in the probit regression.

Table 3. Probit Estimates of Resistant Variety Adoption Equation, Primera and Postrera, 1999-2000

Explanatory Variables [a]	RV adoption (RV=1, TV=0)
Constant	-4.09 (-2.79)**
Season = Postrera	-0.055 (-1.20)
Year = 2000	-0.023 (-0.61)
Intercrop	0.044 (0.63)
Disaster (yield < 80 kg/ha)	0.075 (1.01)
log of Rainfall in first month of season (mm)	0.014 (0.28)
Used formula fertilizer (0=no; 1=yes)	-0.108 (-2.12)**
Used urea fertilizer (0=no; 1=yes)	0.104 (2.15)**
Used both formula & urea (0=no; 1=yes)	0.0005 (1.65)*
Village-level Extension: DICTA	0.729 (1.76)
Zamorano	-0.432 (-3.45)**
Dept of Natural Resources	-0.946 (-4.42)**
Catholic Church	-0.042 (-0.19)
BGYMV area	0.258 (1.80)*
Altitude (m)	0.006 (2.08)**
Altitude^2 (m)	-0.0000 (-1.82)*
Ever Planted Resistant Variety (PRV)	0.355 (7.81)**
[a] Does not include 30 village dummies (z stats in parentheses) ** significant at the 0.05 level * significant at the 0.10 level	N = 900 Log Likelihood = -381.47 Pseudo R2 = 0.343 Coefficients calculated as dF / dX

The factors influencing adoption which are of principal interest to this paper are the exclusion restrictions. A more complete analysis of the factors influencing the adoption of RVs are beyond the scope of this paper and are addressed in a forthcoming paper.

7. Estimation results of bean yield function

To estimate the yield function, we pooled the data over the primera and postrera seasons of 1999 and 2000 (Table 4). We assume that many factors affecting yields, such as year, cropping system, village dummies, etc., have the same effect on both TVs and RVs. However, we assume that the productivity of TVs and RVs are differentially affected by factors such as fertilizer use and season. The regressors in the yield function are nearly the same as those used in the probit function, the difference being the inclusion of the inverse Mills' ratios (IMR) and the omission of the exclusion restriction variables from the probit function.

As expected, the IMR coefficient for TVs is positive, significant at the 0.05 level and of considerable magnitude, thus indicating the presence of selection bias. The positive sign on this coefficient indicates positive selection which means that observed TV yields are higher than what would be observed (positive truncation) if a farmer randomly chosen from the whole sample were to plant a TV. The reason for this is that a farmer selected at random may be in a disease-intensive area, and thus would get low yields, whereas such a farmer would not, in reality, plant a TV.

We did not have an *a priori* expectation for the IMR coefficient term for RVs, though it is found to be positive, significant at the 0.10 level and of smaller magnitude to that of the TV. However, because the sign on the IMR term $\lambda_{Ri} = [- f(\psi_i) / F(\psi_i)]$ is negative, while its corresponding coefficient is positive, this implies that negative selection occurs for RV growers. That is, observed RV yields are lower than they would be in the absence of the unobservables which drive RV use. An explanation for this could be that the unobservable BGYMV pressure (which may well be correlated with other unobserved disease and insect pressures to which the

RVs are not resistant) which drives a farmer to experiment with RVs (PRV) tends to lower the RV yields from what they would be in disease-free areas. In addition, we do not observe many RV growers (and, thus, RV yields) in disease-free areas for two main reasons. First, farmers in areas with low to negligible BGYMV pressure will not likely grow the RVs due to their market price discounts. Second, areas above 1,000 m are more remote and thus have less access to markets and extension - both of which represent access to RV seed and fertilizer.

The results of yield function estimation corrected for selection bias compared with uncorrected OLS results are most distinctively different for the RV dummy and the constant (Table 4). The RV dummy is highly significant and of considerable magnitude in the selection corrected estimation, whereas it is insignificant and of negligible magnitude in the uncorrected estimation. The selection-bias correction increases the magnitude and significance of the RV dummy. All other coefficients in the yield model are essentially of the same significance and magnitude. This means that an RV farmer has higher expected yields than a TV farmer with similar characteristics.

Both corrected and uncorrected estimation show that RVs are more responsive to urea fertilizer than TVs, although the *RV*urea* coefficient is not significant at the 10% level in either equation. However, given the relative magnitude of the RV dummy and those for fertilizer response, it is clear that the principal benefit of RV use is disease resistance, not improved response to key inputs. The coefficient on season (postrera) is negative and significant as expected, given that the postrera season is drier, has more disease pressure, and historically experiences results in lower yields than those of the primera. While we would expect rainfall to increase yield, the coefficient on village rainfall (mm) is insignificant, perhaps due to the

presence of village dummies and the nature of the rainfall data (explained above). While intercropping systems can result in higher bean yields, the significant and positive coefficient on *intercrop* is likely driven by two factors: most intercropping occurs in the primera (when yields are higher), and the conversion of intercropped bean area to "effective monocrop" bean area may underestimate the effective area. The inclusion of village dummies improves the explanatory and thus predictive power of the regression as well as the estimation precision of nearly all the coefficients, although our interest in coefficients relates principally to those for the RV dummy, fertilizer use, and the IMR terms.

Table 4. Adjusted and Unadjusted OLS Estimates of Yield Equation, Primera and Postrera, 1999-2000

Explanatory Variables [a]	Corrected OLS ln(yield)	Uncorrected OLS ln(yield)
Constant	5.630 (14.74)**	5.718 (14.82)**
Season = Postrera	- 0.220 (-2.97)**	-0.238 (-3.23)**
Year = 2000	- 0.192 (-2.85)**	-0.193 (-2.85)**
Intercrop	0.245 (2.52)**	0.269 (2.81)**
Disaster (yield < 80 kg/ha)	-3.369 (-14.80)**	-3.36 (-14.69)**
log of Rainfall in first month of season (mm)	0.021 (0.29)	0.026 (0.35)
Used formula fertilizer (0=no; 1=yes)	0.206 (2.14)**	0.151 (1.64)
Used urea fertilizer (0=no; 1=yes)	0.060 (0.76)	0.098 (1.27)
Used both formula & urea (0=no; 1=yes)	- 0.0004 (-1.53)	- 0.0005 (-1.46)
RV	0.552 (2.73)**	0.129 (1.10)
RV * Season = Postrera	- 0.175 (-1.52)	- 0.177 (-1.56)
RV * Year = 2000	- 0.020 (-0.18)	- 0.015 (-0.14)
RV * Disaster	- 0.554 (-1.48)	- 0.554 (-1.47)
RV * Formula Fertilizer	- 0.025 (-0.16)	- 0.013 (-0.08)
RV * Urea Fertilizer	0.155 (1.09)	0.151 (1.07)
RV * Both Formula & Urea Fertilizers	-0.0006 (-0.51)	-0.0007 (-0.54)
Inverse Mills' Ratio (TVs)	0.342 (2.51)**	–
Inverse Mills' Ratio (RVs)	0.236 (1.69)*	–
[a] Does not include 30 village dummies (t stats in parentheses) ** significant at the 0.05 level * significant at the 0.10 level	N = 900 R2 = 0.680 F(46, 854) = 16.07	N = 900 R2 = 0.678 F(44, 856) = 16.36

8. Counterfactual Predictions

Estimation of the selection corrected yield function enables the prediction of imputed yields for each individual's unobserved varietal choice, evaluating each individual's own

observed characteristics at the coefficients from the alternate subsample. Finally, the predicted TV yields of RV users can be statistically compared with predictions of RV yields to estimate the farm-level impact of disease-resistant bean varieties in Honduras.

The significant coefficient on the *RV* dummy imply that the expected yield of RV users is higher than that of TV users, *ceteris paribus*. However, the more appropriate counterfactual scenario for impact analysis is the predicted yield that RV users would have obtained in the absence of RVs – the TV yield of an RV grower. The conditional counterfactual will produce the yield differential between predicted RV and TV yields for a farmer chosen at random from the RV subsample. This conditional counterfactual[8] is computed for RV growers as the selection corrected constant (common to both RV and TV growers) plus the relevant IMR coefficient multiplied by the corresponding IMR term, plus the characteristics of RV growers multiplied by the TV coefficients estimated from the TV subsample.

After obtaining predictions from each regression, the mean values of log(yield) are

[8] The conditional counterfactual assumes that the grower has already selected into the RV subsample and includes the IMR term. Unconditional counterfactual would not assume that the grower had already selected into the RV subsample, and would not include the IMR term in the computation of predictions. While predicted yield levels vary given which counterfactual is computed (Table 5b), the percentage differential between the predicted RV yield and the predicted RV counterfactual is not affected by this choice. The reason for this is that since these differentials are calculated using predictions from the same subsample, it doesn't matter whether or not the IMR term is included. However, if we were to compute a differential between predicted RV yields (for RV users) and predicted TV yields (for TV users) – a differential across subsamples – then the counterfactual choice makes a large difference. For example, the differential between conditional predictions across subsamples are almost identical to those from OLS (605 vs. 541 kg/ha is 11%), while differentials between unconditional predictions are much larger (714 vs 476 kg/ha is 33%). Conditional differentials would thus be expected to be smaller than unconditional estimates, in which the varietal choice has not yet been made.

Table 5. Predictions of RV and TV bean yields from Corrected and Uncorrected Yield Models, 1999-2000

Comparison of Mean Predictions by Subsample	Sub-Sample Size	Corrected OLS Estimation*				Uncorrected OLS Estimation			
		Mean Predicted Yields (kg / ha)		Differential [b] (%)		Mean Prediction (kg / ha)		Differential (%)	
RV vs TVcf $\mid_{RV}$ [a]	N=312	605	361	40	35	634	577	9	0
RV vs TVcf $\mid_{RV}$: Primera	N=145	757	418	45	42	792	666	16	12
RV vs TVcf $\mid_{RV}$: Postrera	N=167	473	313	34	28	496	501	- 1	- 9
TV vs RVcf $\mid_{TV}$ #1(adds RV, RV*fert)	N=588	541	889	- 64	- 56	568	612	- 8	- 3
TV vs RVcf $\mid_{TV}$ #2 (adds RV)	N=588	541	860	- 59	- 52	568	591	- 4	0
TV vs RVcf $\mid_{TV}$ #3 (adds RV*fert)	N=588	541	512	5	9	568	538	5	10
TV vs RVcf $\mid_{TV}$ #3 : Primera	N=245	622	641	- 3	0	652	673	- 3	0
TV vs RVcf $\mid_{TV}$ #3 : Postrera	N=343	483	421	13	17	507	441	13	17
TV vs RVcf $\mid_{TV}$ #4 (adds neither RV nor RV*fert)	N=588	541	495	9	12	568	519	9	12

* Corrected predictions include IMR coefficients and IMR terms
Observed mean sample yields: RV = 673 kg/ha; RV primera = 810 kg/ha; RV postrera = 555 kg/ha
Observed mean sample yields: TV = 582 kg/ha; TV primera = 654 kg/ha; TV postrera = 531 kg/ha

[a] cf = counterfactual
[b] Left differential is calculated as the ratio of the difference between sample mean yield predictions (from the mean predicted yields columns) divided by the first term; for eg., [(RV - TVcf $\mid_{RV}$) / RV] * 100. Right differential is calculated as the mean of farmer differentials, calculated by the same ratio, but for each farmer; all left differentials are significant at the 0.01 level except where otherwise indicated.

scaled to levels[9]. The mean predictions (Table 5) are similar to actual means, although the predicted RV yields are lower in general than actual RV yields. Two differentials are presented (Table 5); the first is calculated as the ratio of the difference between sample mean yield predictions (from the mean predicted yields columns) divided by the first term; for eg., [(RV - $TVcf|_{RV}$) / RV] * 100; while the second is calculated as the mean of farmer differentials, calculated by the same ratio, but for each farmer.

Given that we found negative selection for RV yields (mean RV yields are lower than they would be if RVs were grown in areas of no disease pressure) and positive selection for TV yields (TV yields are higher than they would be if TVs were grown in areas of high disease pressure), we would expect that differentials calculated from the selection-corrected yield equation would be larger than those from the uncorrected OLS equation. As expected, comparison of differentials by estimation technique shows that those from the corrected model are significantly larger than those from the uncorrected model (Table 5); the corrected model predicts the mean farmer differential to be 42% in the primera, and 28% in the postrera, while the uncorrected model predicts a differential of 12% and -9%, respectively for these two seasons.

Furthermore, due to market price differentials by variety, yield differentials must be adjusted in order to more accurately reflect the net gain to farmers of varietal choice. For example, the valuation of an RV grower's counterfactual (using a TV) must account for the fact

[9] The conversion of predicted log yield to level yield is exp(*scale*) * exp(ln(*yield_hat*)) where *scale* is computed as the coefficient B_0 of the regression *Yield_hat* = B_0 exp(ln(*yield_hat*)) with no constant.

that while the RV may yield more than the TV, the market price of an RV is 15% less relative to a TV[10]. Even after adjusting the differentials for market price discounts, the corrected model predicts that RV growers enjoy net income gains of 27% in the primera and 13% in the postrera compared to what they would have earned growing TVs.

However, adjusting the uncorrected model differentials yields counterintuitive results: farmers growing RVs would lose 3% in net income in the primera and lose 24% in net income in the postrera relative to what they would have earned growing TVs. This result implies that RV growers do not enjoy any yield gain (yield loss averted) to compensate for the price discount (and actually lose net income in both seasons), yet continue to grow RV nevertheless. By contrast, RV differentials from the corrected model more than offset the price discount in both seasons. The conflicting results from these two models highlight the sensitivity of impact analysis of maintenance technologies to the method of econometric measurement of farm-level net benefits: using farm-level net benefits (differentials) from the corrected model yields positive aggregate benefits, whereas using differentials from the uncorrected model yields negative aggregate benefits.

Counterfactuals are also computed for TV growers as a means of testing the model, wherein we compute a predicted TV yield and add the RV coefficient and the RV*fertilizer coefficients (for those TV farmers who use fertilizers). While we would expect that predicted differentials for TV growers would be positive, or if negative, would not be larger than the price

[10] Dorado and Don Silvio, released in the early 1990s, represent 75% of current RV users. The mean price discount for these varieties is 15%. However, Tio Canela, released in 1997 and representing 25% of RV users, has a smaller farm-level price discount of 7% (Mather *et al*, 2003).

discount, the model in fact predicts that TV growers would see sizeable net gains from switching to RVs. As indicated by Table 5, these results are driven by the RV coefficient which is added to the TV counterfactual. Given that access to RVs is not likely a critical factor in the adoption decision, this result can perhaps be explained by considering that the RVs were targeted for valley farmers, whose growing conditions may be quite different from higher-altitude farmers. That is, the RV dummy represents a yield effect enjoyed by RV adopters in disease-prone areas (which may well have other disease and pest pressures beyond simply BGYMV) over and above BGYMV resistance, that the average TV grower will not enjoy. If this assumption is correct, then the more appropriate counterfactual for TV growers is #3 (Table 5), wherein we compare the mean predicted TV yield with the mean predicted TV yield which includes the fertilizer effects of RV growers. These counterfactual values appear more reasonable as predicted RV yields are not high enough to offset the RV price discount. Another explanation for the large TV differentials could be that the model consists primarily of dummy variables, and thus we may not have enough variation among the characteristics (regressors) of TV farmers to capture the differential response of TV and RV varieties to different farmer input quantities and qualities.

9. Implications for Impact Analysis

Using the economic surplus model and costs of research as outlined in paper #1, rates of return to disease-resistant bean research in Honduras are calculated for the postrera seasons 1982-2010 (Table 6), using differentials from the corrected and uncorrected models, as well as the differential derived from an expected utility framework (paper #1; Mather *et al*, 2003). The results demonstrate the significant implications of the method described in this paper for the construction of counterfactual scenarios for use in the analysis of the impact of maintenance

technologies. Use of uncorrected OLS results would lead an analyst to conclude that RV use generated negligible or negative returns, while the corrected OLS results generate a high ROR. The expected utility framework provides a minimum estimate of the net farmlevel benefits of RV use, which appear to have underestimated the benefits in comparison with the corrected OLS approach using sample mean differentials.

Table 6. Rates of Return to Disease-Resistant Bean Research in Honduras, Postrera Seasons 1982-2010

Method of Calculating Incremental Farm-level net benefits of RV use		Incremental net farmlevel benefit of Dorado use in the postrera season, adjusted for price discount (%)	Economic Rate of Return [a] (%)
Expected Utility Framework [b]		11	41.2
Farmer Mean Differential	Corrected Model:	13	45.0
	Uncorrected Model:	- 24	–
Sample Mean Differential	Corrected Model:	19	54.5
	Uncorrected Model:	- 16	–

[a] Calculations use the economic surplus model as outlined in first essay.
[b] Framework and benefits as presented in first essay.

10. Conclusions

This paper demonstrates a method for using farmlevel survey data in the construction of counterfactual scenarios for use in impact assessment of maintenance research. The method uses a modification of Lee's (1978) two-step procedure to correct for selection bias, the presence of which in farmlevel survey data will likely lead to underestimation of the benefits of maintenance research (Johnson *et al*, 1999). The method is applied to test for selection bias and estimate yield differentials between RV and TV bean yields in Honduras by constructing the counterfactual to RV yields as what yields RV users would have obtained had they continued to

grow TVs (i.e. the TV yield of RV users). The corrected yield model predicts that RV growers enjoy net income gains of 13 to 19 percent compared to what they would have earned growing TVs, while the uncorrected OLS model predicts that RV growers see either no income gain or even losses by growing RVs (relative to TVs). This application demonstrates that the implications of this method are significant for the assessment of maintenance research impacts, both at the farm-level and in terms of the ROR to research investments.

References

Duncan, G., D. Leigh. 1980. Wage Determination in the Union and Nonunion Sectors: A Sample Selectivity Approach. Industrial and Labor Relations Review 34:24-34.

Goldfeld, S.M., Quandt, R.E. 1973. The Estimation of Structural Shifts by Switching Regressions. Annals of Economic and Social Measurement 2:475-485.

Gronau, R. 1974. Wage Comparisons – A Selectivity Bias. Journal of Political Economy 82:1119-43.

Heckman, J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models. Annals of Economic and Social Measurement 5:475-492.

Heckman, J. 1979. Sample Selection Bias as a Specification Error. Econometrica 47:153-161.

Johnson, N., Klass, J., 1999. The Impact of Crop Improvement Research on Rural Poverty: A Spatial Analysis of BGYMV-Resistant Bean Varieties in Honduras. Paper prepared for the workshop: Assessing the Impact of Agricultural Research on Poverty Alleviation, San Jose, Costa Rica, September 14-16, 1999.

Lee, L.F. 1976. Estimation of Limited Dependent Variable Models by Two-Stage Methods. Ph.D. dissertation. University of Rochester.

Lee, L.F. 1978. Unionism and wage rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables. International Economic Review 19:415-433.

Lewis, H.G. 1974. Comments on Selectivity Biases in Wage Comparisons. Journal of Political Economy 82(6):1145-55.

Mather, D., Bernsten, R., Rosas, J.C., Viana, A.R., Escoto, D. 2003 (forthcoming). The Economic Impact of Disease-Resistant Bean Research in Honduras. Agricultural Economics.

Martel, P., Bernsten, R., Weber, M., 2000. Food Markets, Technology, and the Case of Honduran Dry Beans. Michigan State University International Development Papers, Working Paper No.78. Department of Agricultural Economics, Michigan State University, East Lansing, MI.

Martel, P., 1995. A Socio-Economic Study of the Honduran Bean Subsector: Production Characteristics, Adoption of Improved Varieties, and Policy Implications. Ph.D. Dissertation, Department of Agricultural Economics, Michigan State University, East Lansing, MI.

Otsuka, K., F. Gascon, S. Asano. 1994. Second-generation MVs and the evolution of the Green Revolution: the case of Central Luzon, 1966-90. Agricultural Economics 10:283-295.

Roy, A.D. 1951. Some Thoughts on the Distribution of Earnings. Oxford Economic Papers 3:155-146.

Smale, M., Singh, R.P., Sayre, K., Pingali, P., Rajaram, S., Dubin, H.J., 1998. Estimating the economic impact of breeding nonspecific resistance to leaf rust in modern bread wheats. Plant Dis. 82:1005-61.

Vella, F. 1998. Estimating Models with Sample Selection Bias: A Survey. The Journal of Human Resources. 32:127-169.

Wooldridge, J. 2000. Introductory Econometrics: A Modern Approach. South-Western College Publishing.

Table 5b. Predictions of RV and TV bean yields from Unconditional and Conditional Corrected Yield Models, 1999-2000

Comparison of Mean Predictions by Subsample	Sub-Sample Size	Corrected OLS Estimation without IMR				Corrected OLS Estimation with IMR			
		Mean Predicted Yields (kg / ha)		Differential [b] (%)		Mean Prediction (kg / ha)		Differential (%)	
RV vs TVcf $\mid_{RV}$ [a]	N=312	714	427	40	35	605	361	40	35
RV vs TVcf $\mid_{RV}$: Primera	N=145	882	487	45	42	757	418	45	42
RV vs TVcf $\mid_{RV}$: Postrera	N=167	569	376	34	28	473	313	34	28
TV vs RVcf $\mid_{TV}$ #1(adds RV, RV*fert)	N=588	476	782	- 64	- 56	541	889	- 64	- 56
TV vs RVcf $\mid_{TV}$ #2 (adds RV)	N=588	476	757	- 59	- 52	541	860	- 59	- 52
TV vs RVcf $\mid_{TV}$ #3 (adds RV*fert)	N=588	476	450	5	9	541	512	5	9
TV vs RVcf $\mid_{TV}$ #3 : Primera	N=245	546	562	- 3	0	622	641	- 3	0
TV vs RVcf $\mid_{TV}$ #3 : Postrera	N=343	427	371	13	17	483	421	13	17
TV vs RVcf $\mid_{TV}$ #4 (adds neither RV nor RV*fert)	N=588	476	436	9	12	541	495	9	12

Observed mean sample yields: RV = 673 kg/ha; RV primera = 810 kg/ha; RV postrera = 555 kg/ha
Observed mean sample yields: TV = 582 kg/ha; TV primera = 654 kg/ha; TV postrera = 531 kg/ha

[a] cf = counterfactual
[b] Left differential is calculated as the ratio of the difference between sample mean yield predictions (from the mean predicted yields columns) divided by the first term; for eg., [(RV - TVcf $\mid_{RV}$) / RV] * 100. Right differential is calculated as the mean of farmer differentials, calculated by the same ratio, but for each farmer; all left differentials are significant at the 0.01 level except where otherwise indicated.
[c] significant at 0.05 level [d] insignificant at 0.10 level