ON THE LINEAR REGRESSION OF TRUE OBSERVATIONS WITH CLASSIFIER DERIVED ESTIMATES

Richard P. Heydorn, NASA Johnson Space Center

ABSTRACT: A crop sample survey method which regresses ground gathered acreage determinations with computer classifications of remotely sensed satellite data is considered. The resulting regression functions are related to the characteristics of the classifier. In particular, if y is an observed proportion of a given crop in a group of sampling units and x is the classifier derived estimate of that proportion, then the regression function values $E(X|Y=y)$ and $E(Y|X=x)$ are derived in terms of conditional probabilities of certain Bernoulli events defined on the sampling units. As a consequence, this indicates that classifier designs which are able to keep these probabilities constant across certain groupings of sampling units insure that the regression functions will be linear.

1.0 INTRODUCTION: Sample survey designs have been proposed in which sampled observations are regressed onto the outcomes of some auxiliary variable to increase the precision of the survey. Generally, the higher the correlation between the survey variable and the auxiliary variable, the lower is the variance of the regression derived estimates. One such design is being studied by the USDA (1). In their approach, estimates of crop acreages are acquired by trained enumerators using a sampling approach. The estimates are then linearly regressed onto estimates acquired from a satellite over the same sampling units. Once the regression estimator is formed, satellite estimates obtained at a much higher sampling rate are transformed by the regression estimator to obtain the crop area estimate. In this process the satellite acquired data for a given resolution element (pixel) on the ground is dichotomously classified into the crop of interest or the class of "all other." Classifications of this kind are averaged over sampling units to obtain the satellite derived acreage estimates.

In this paper we examine the properties of the classifier that are needed to satisfy the conditions of a linear regression model in this sample survey setting.

2.0 PROBLEM DEFINITION: Let Ω be the set of all outcomes related to the selection of a given sampling unit and let V be the set of all outcomes related to the selection of a pixel. Thus, pixel $v \in V$ from sampling unit $w \in \Omega$ will be denoted by (v,w). Let (Z,Θ) be a random variable of pixel observations where $Z(v,w) \in \mathcal{R}^N$ ($\mathcal{R}^N$ a real product spare) and $\Theta(v,w) \in \{0,1\}$. The variable Z denotes the satellite measurements that are to be classified and Θ denotes the true crop label of a pixel. If $\Theta(v,w)=1$, then the crop of interest is being observed and if $\Theta(v,w)=0$ then some other crop is under observation. Values for Θ will only be known for the sampling units in the ground enumerative survey. In the sequence to follow, we will also have occasion to consider a random variable ∇ on Ω where $\nabla(w)=w$.

Let $\phi: \mathcal{R}^N \to \{0,1\}$ be a classifier where ϕ is derived so that, by some criterion, $\phi(Z(v,w))$ is close to $\Theta(v,w)$. The classifications can be thought of as extensions into sampling units where no ground observations have been made. A discussion of how such a classifier can be derived will not be given. It suffices to say, however, that in the usual setting a parametric form is first postulated and then the parameter values are estimated using a portion of the ground acquired observations. This process is often called "training the classifier." The strategy used to train a classifier as well as the number of training samples used can introduce varying amounts of variance in the final acreage estimates. In this development we shall avoid these added complexities by assuming that the classifier is given, that is, it has not been derived through a training process.

Denote the means of sampling unit "w" as $Y(w) = E(\Theta(w))$ and $X(w) = E(\phi(Z(w))$ and consider the models

$$Y = a + bX + \varepsilon \tag{1}$$

and

$$X = c + dY + \eta \tag{2}$$

These models will be linear if, and only if, (iff) the error terms are conditionally unbiased estimates of 0. That is equation (1) will be linear iff $E(\varepsilon|X)=0$ and similarly equation (2) will be linear iff $E(\eta|Y)=0$. The properties of $\phi(Z)$, relative to Θ, that will guarantee that these models are linear is the main subject of this paper.

3.0 THE TWO-POINT MODEL: Given a sampling unit, w, consider the models

$$\Theta = \alpha_w + \beta_w\phi(Z) + \xi \tag{3}$$

$$\phi(Z) = \gamma_w + \delta_w\Theta + \zeta \tag{4}$$

where the subscript w indicates the fact that the constants are sampling unit dependent. These constants are defined to be

$$\alpha_w = \Pr(\Theta=1|\phi(Z)=0, \Delta=w)$$
$$\beta_w = \Pr(\Theta=1|\phi(Z)=1, \Delta=w) - \alpha_w$$
$$\gamma_w = \Pr(\phi(Z)=1|\Theta=0, \Delta=w)$$
$$\delta_w = \Pr(\phi(Z)=1|\Theta=1, \Delta=w) - \gamma_w$$

It is easy to see that these coefficients are the least square coefficients and that $E(\xi|\phi(Z)) = E(\xi|\Theta)=0$; i.e., the models are linear.

The regression functions $E(Y|X)$ and $E(X|Y)$ will now be derived in terms of these two-point models.

4.0 THE MAIN THEOREMS: We first define the sets.

$$A_x = \{w:X(w) = x\}$$
$$B_y = \{w:Y(w) = y\}$$

where we will assume that A_x and B_y are nonempty for any x,y in [0,1].

Theorem 1

$$E(Y|X) = \bar{\alpha}_X + \bar{\beta}_X X$$
$$E(X|Y) = \bar{\gamma}_Y + \bar{\delta}_Y Y$$

where

$$\bar{\alpha}_X = \Pr(\Theta=1|\phi(Z) = 0, \ \Delta\varepsilon A_X)$$
$$\bar{\beta}_X = \Pr(\Theta=1|\phi(Z) = 1, \ \Delta\varepsilon A_X) - \bar{\alpha}_X$$
$$\bar{\gamma}_Y = \Pr(\phi(Z)=1|\Theta=0, \ \Delta\varepsilon B_Y)$$
$$\bar{\delta}_Y = \Pr(\phi(Z)=1|\Theta=1, \ \Delta\varepsilon B_Y) - \bar{\gamma}_Y.$$

Proof: We prove the first assertion; the second follows by a similar proof. From equation (3)

$$E(\Theta|\phi(Z), \Delta=w) = \alpha_w + \beta_w\phi(Z)$$

Therefore

$$E(E(\Theta|\phi(Z), \Delta=w)|\phi(Z), \Delta\varepsilon A_X) = E(\alpha_w|\phi(Z), \Delta\varepsilon A_X) + E(\beta_w|\phi Z), \Delta\varepsilon A_X)\phi(Z)$$

or

$$E(\Theta|\phi(Z), \Delta\varepsilon A_X) = (E(\alpha_w|\phi(Z), \Delta\varepsilon A_X) + E(\beta_w|\phi(Z), \Delta\varepsilon A_X)\phi(Z).$$

Now

$$E(\Theta|\phi(Z) = 0, \Delta\varepsilon A_X) = E(\alpha_w|\phi(Z) = 0, \Delta\varepsilon A_X)$$

which implies

$$E(\alpha_w|\phi(Z) = 0, \Delta\varepsilon A_X) = \Pr(\Theta=1|\phi(Z) = 0, \Delta\varepsilon A_X) \triangleq \overline{\alpha}_X$$

Also

$$E(\Theta|\phi(Z) = 1, \Delta\varepsilon A_X) = E(\alpha_w + \beta_w|\phi(Z) = 1, \Delta\varepsilon A_X)$$

or

$$E(\alpha_w + \beta_w|\phi(Z) = 1, \Delta\varepsilon A_X) = \Pr(\Theta=1|\phi(Z) = 1, \Delta\varepsilon A_X) \triangleq \overline{\beta}_X + \overline{\alpha}_X$$

Hence

$$E(\Theta|\phi(Z), \Delta\varepsilon A_X) = \overline{\alpha}_X + \overline{\beta}_X\phi(Z)$$

Let (Z_i,Θ_i) i=1,2,... be an iid sample from A_X. Then for any i we have from independence

$$E(\Theta_i|\alpha(Z_i), \Delta\varepsilon A_X) = E(\Theta_i|\phi(Z_1),\phi(Z_2),..., \Delta\varepsilon A_X)$$

Hence, for a given n

$$E(\frac{1}{n}\sum_{i=1}^{n}\Theta_i|\alpha(Z_1),\alpha(Z_2),..., \Delta\varepsilon A_X) = \overline{\alpha}_X + \overline{\beta}_X \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i)$$

Since, for the given sample (Z_i, Θ_i), i=1,2,...

$$\frac{1}{n}\sum_{i=1}^{n}\Theta_i \xrightarrow{a.s.} E(Y|\Delta\varepsilon A_X) = E(Y|X)$$

$$\frac{1}{n}\sum_{i=1}^{n}\phi(Z_i) \xrightarrow{a.s.} E(X|\Delta\varepsilon A_X) = X$$

and since conditional expectation is a bounded linear operator, we have

$$E(E(Y|X)|\phi(Z_1), \phi(Z_2),..., \Delta\varepsilon A_X) = \overline{\alpha}_X + \overline{\beta}_X X$$

Moreover, since events $\mathcal{B}(\frac{1}{n}\sum_{i=1}^{n}\phi(Z_i))\subset\mathcal{B}(\phi(Z_1), \phi(Z_2),...)$* for all n, we have upon conditioning on $\mathcal{B}(X)$

$$E(E(Y|X)|X) = \overline{\alpha}_X + \overline{\beta}_X X$$

or

$$E(Y|X) = \overline{\alpha}_X + \overline{\beta}_X X$$

which completes the proof.

Theorem 1 gives sufficient conditions which will guarantee that the regression functions E(Y|X) and E(X|Y) be linear; viz, that the corresponding functions $\overline{\alpha}_X$, $\overline{\beta}_X$, $\overline{\gamma}_Y$, and $\overline{\delta}_Y$ be constants.**

In the case of E(X|Y) this leads to an interesting interpretation of the role of the classifier. Consider the family $F = F_0 \cup F_1$ where

$$F_0 = \{F_0(\cdot;y):y\varepsilon[0,1]\}$$
$$F_1 = \{F_1(\cdot;y):y\varepsilon[0,1]\}$$

where

$$F_i(z;y) = \Pr(Z<z|\Theta=i, \Delta\varepsilon B_y), \quad i=0,1$$

Thus, if ϕ is an ancillary statistic on F, E(X|Y) is a linear regression function. This means that the classifier should smooth over sampling unit differentes--i.e., if a classifier design were possible that could make the omission and commission errors*** constant at least over groups B_y, then the regression of X on Y would be linear. In the model $X = c + dY + \eta$ this ancillary condition therefore implies that η will be a conditionally unbiased estimate of 0.

If we randomly draw from the collection of all sampling units, rather than from just one, as was done in equation (4), then we again have a two-point model of the form

$$\phi(Z) = c + d\Theta + \zeta, \quad E(\zeta|\Theta) = 0$$

where

$$c = \Pr(\phi(Z) = 1|\Theta = 0)$$
$$d = \Pr(\phi(Z) = 1|\Theta = 1) - c$$

Let (Z_i, Θ_i) i=1,2,...,n be a sample of size n from a particular sampling unit and define the averages

$$X_n = \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i) \qquad (5)$$

$$Y_n = \frac{1}{n}\sum_{i=1}^{n}\Theta_i \qquad (6)$$

If we find that all such averages have the same linear regression, viz,

$$X_n = c + dY_n + \zeta_n, \quad E(\zeta_n|Y_n) = 0$$

then this condition completely characterizes the ancillary property of the classifier. We state this more formally in the following theorem.

Theorem 2

ϕ is ancillary for the family F iff there exists constants c,d such that $E(X_n|Y_n) = c + dY_n$ for all n.

Proof: Assume $E(X_n|Y_n) = c + dY_n$.

We have for any n,N, N>n

$$E(\zeta_n|Y_n,Y_N) = E(\zeta_n|nY_n, NY_N - nY_n)$$

But ζ_n depends only on sums up to n and not beyond so that

$$E(\zeta_n|Y_n,Y_N) = E(\zeta_n|Y_n)$$

and since this holds for all N and $Y_N \xrightarrow{a.s.} Y$ we have,

$$E(\zeta_n|Y_n,Y) = E(\zeta_n|Y_n)$$

Thus, since $E(\zeta_n|Y_n) = 0$, we have

$$E(Y_n\zeta_n|Y) = E(E(Y_n\zeta_n|Y_n,Y)|Y) = E(Y_nE(\zeta_n|Y_n,Y)|Y) = E(Y_nE(\zeta_n|Y_n)|Y) = 0 \qquad (7)$$

We have the model

$$X_n = c + dY_n + \zeta_n$$

and therefore, using equation (7), and the fact that $E(Y_n|Y) = Y$,

$$Y_nX_n = cY_n + dY_n^2 + Y_n\zeta_n$$
$$E(Y_nX_n|Y) = cY + dE(Y_n^2|Y)$$

or

$$E((Y_n-Y)X_n|Y) + YE(X_n|Y) - cY - dY^2 = dE((Y_n-Y)^2|Y)$$

Noting that $E(X_n|Y) = E(X|Y)$ and $E((Y_n-Y)E(X_n|Y)|Y) = 0$, we have

$$E((Y_n-Y)(X_n-E(X_n|Y))|Y) + Y(E(X|Y) - c - dY) = dE(Y_n-Y)^2|Y)$$

and since $E(\zeta_n|Y) = E(E(\zeta_n|Y_n)|Y) = 0$ for all n implies $E(X|Y) - c - dY = 0$, we have

$$\frac{\mathrm{cov}(Y_n,X_n|Y)}{\mathrm{var}(Y_n|Y)} = d$$

Now

$$\frac{\mathrm{cov}(Y_n,X_n|Y)}{\mathrm{var}(Y_n|Y)} = \frac{E(\Theta\phi(Z)|Y)-E(\Theta|Y)E(\phi(Z)|Y)}{E(\Theta|Y)-E(\Theta|Y)^2}$$

and after some manipulation this reduces to

$$d = \Pr(\phi(Z)=1|\Theta=1, \Delta\varepsilon B_Y) - \Pr(\phi(Z)= 1|\Theta=0, \Delta\varepsilon B_Y) = \overline{\delta}_Y$$

---

*Given the random variable V, the notation $\mathcal{B}(V)$ means the sigma field generated by V.

**We recognize that other conditions will guarantee linearity, e.g., $\overline{\alpha}_X$, $\overline{\gamma}_Y$ linear and $\overline{\beta}_X$, $\overline{\delta}_Y$ constants, but these conditions are not discussed.

***Within a B_y-group commission error is defined as $\Pr(\phi(Z)=0|\Theta=1, \Delta\varepsilon B_y)$ and the omission error is defined as $\Pr(\phi(Z)=0|\Theta=1, \Delta\varepsilon B_y)$.

By Theorem 1, $E(X|Y) = \bar{\gamma}_Y + \bar{\delta}_Y$ and thus

$$E(X|Y) = \bar{\gamma}_Y + dY$$

but, by the assumption that $E(X|Y) = c + dY$ this means that

$$\bar{\gamma}_Y = c$$

Hence, both $\Pr(\phi(Z)=1|\Theta=1, \Delta\varepsilon B_Y)$ and $\Pr(\phi(Z)=0|\Theta=0, \Delta\varepsilon B_Y)$ are independent of Y and so ϕ is ancillary for F.

To prove the assertion the other way, we have proceeding as we did in the proof of Theorem 1.

$$E(X_n|Y_n, \Delta\varepsilon B_Y) = \bar{\gamma}_Y + \bar{\delta}_Y Y_n$$

But, since ϕ is ancillary for F, we have $\bar{\gamma}_Y = c$, $\bar{\delta}_Y = d$ and therefore

$$E(X_n|Y_n) = E(E(X_n|Y_n, \Delta\varepsilon\beta_Y)|Y_n) = c + dY_n$$

which completes the proof.

A similar theorem holds when the regression function $E(Y|X)$ is considered.

Theorem 3

$\bar{\alpha}_X$ and $\bar{\beta}_X$ are independent of X iff there exists constants a, b such that $E(Y_n|X_n) = a + bX_n$ for all n.

The proof of this theorem is analogous to that of Theorem 2.

5.0 CLASSIFIER DESIGN: Theorem 1 shows that the regression functions $E(Y|X)$ and $E(X|Y)$ are determined respectively by the functions $\bar{\alpha}_X$, $\bar{\beta}_X$ and $\bar{\gamma}_Y$, $\bar{\delta}_Y$. Since these functions are related to the responses of the classifier, they provide a means for understanding the effect of the classifier on a linear regression, or any other regression form. If the regression function $E(X|Y)$ is being considered then, as was pointed out above, a classifier that is an ancillary statistic is a desirable classifier to insure that $E(X|Y)$ be linear.

If a linear functional form for $E(Y|X)$ is desired, then a property somewhat opposite to ancillary is needed if $\bar{\alpha}_X$ and $\bar{\beta}_X$ are to be constant. This property is analogous to the property of a sufficient statistic. However the connection with a sufficient statistic is not as clear as is the previously discussed connection in the case of the ancillary statistic. Part of the problem is that the conditional probabilities $\Pr(\Theta=1|\phi(Z) = 0, \Delta\varepsilon A_X)$ and $\Pr(\Theta=1|\phi(Z) = 1, \Delta\varepsilon A_X)$, which define $\bar{\alpha}_X$ and $\bar{\beta}_X$, are indexed on x-values rather than y-values. Nevertheless, the role of the classifier in this formulation is essentially one of preserving information regarding sampling unit crop proportions as opposed to smoothing over that information as is done in the case of the ancillary statistic.

If the classifier should have properties that tend to preserve the crop proportion information in a sampling unit, then it would be interesting to examine $E(Y|X)$ when a Bayes classifier is used and when a classifier such as a maximum likelihood classifier is used. Since the Bayes classifier uses the actual proportion of the crop of interest as a prior probability, one may conjecture that such a classifier would keep $\bar{\alpha}_X$ and $\bar{\beta}_X$ reasonably constant over x-values whereas the maximum likelihood classifier, which ignores these priors, would not.

To gain some insight into this question, a simple example was constructed in which Z for the crop of interest in a given sampling unit has a normal probability density $N(\cdot;\mu_1,\sigma)$ and the class of all other material is distributed as $N(\cdot;\mu_0,\sigma)$. Sampling units were simulated by treating μ_0 and y as independent uniformally distributed values. The ranges on μ_0 and y were from .1 to .9. The other parameters μ_1 and σ were held fixed at values of 0 and .25 respectively.

In a given sampling unit where y is the proportion of the crop of interest and z represents the spectral value of a pixel, the Bayes classifier decides the pixel is the crop of interest if

$$yN(z;\mu_1,\sigma) > (1-y)\, N(z;\mu_0,\sigma)$$

otherwise the pixel is classified into the class of "all other." For the maximum likelihood classifier the decision strategy is similar except that the priors (y-values) are ignored. This is the same as letting y=1/2 in the Bayes classifier.

Figures 1a to 1f show the resulting $\bar{\alpha}_X$ and $\bar{\beta}_X$ values for the two classifiers. The Bayes classifier produces a regression function that is very nearly linear in the shown domain of x-values. While $\bar{\beta}_X$-values are nearly constant, the $\bar{\alpha}_X$-values have a slight positive slope indicating that a part of the slope of $E(Y|X)$ comes from the $\bar{\alpha}_X$ function. The maximum likelihood classifier produces a different regression function. It is more "S"-shaped and the corresponding $\bar{\alpha}_X$ and $\bar{\beta}_X$ functions are also nonlinear. Thus, if the objective is to produce a linear regression, then the maximum likelihood classifier is not a satisfactory choice, at least for this case. The Bayes classifier is better, but it falls somewhat short of holding $\bar{\alpha}_X$ constant (assuming constant $\bar{\alpha}_X$-values are desirable).

The regression function from the maximum likelihood classifier tends to have large curvature values, in this example, for small and large x-values. The reason is that the maximum likelihood classifier tends to overestimate when the crop of interest exists in small proportion and tends to underestimate when the proportion is large. The overestimation is due to the commission error rate of the classifier and the underestimation is due to omission errors. The Bayes classifier on the other hand, tries to adjust itself so as to reduce these commission and omission errors at the extremes of the proportion values. Indeed, if the prior probability is low for the crop of interest, then the Bayes classifier will assign more of the pixels to the opposite class than will the maximum likelihood classifier with the result of reducing the commission error. In high proportion situations the opposite is true.

6.0 SAMPLING EFFICIENCY: Consider the linear model given by equation (1) and let $a = E(Y) - bE(X)$, then

$$E(Y) = Y + b(E(X)-X) - \varepsilon$$

and let

$$\widehat{E(Y)} = Y + b(E(X)-X)$$

be our estimator of crop acreage. (This is essentially the formulation given by Cochran (2) page 189). When b is the usual least square coefficient, the variance of this estimator is easily seen to be

$$\mathrm{Var}(\widehat{E(Y)}) = \mathrm{Var}(Y)\,(1-R^2)$$

where

$$R = \frac{\mathrm{COV}(X,Y)}{\mathrm{var}(X)\ \mathrm{var}(Y)} = b\,\frac{\mathrm{var}(X)}{\mathrm{var}(Y)}$$

For the case where $\bar{\beta}_X = b$ for all x, we have $b = \Pr(\Theta=1|\phi(Z)=1) - \Pr(\Theta=1|\phi(Z)=0)$ which can be rewritten as

$$b = \Pr(\phi(Z)=1|\Theta=1)\frac{\Pr(\Theta=1)}{\Pr(\phi(Z)=1)} - \Pr(\phi(Z)=0)|\Theta=1)\ \frac{\Pr(\Theta=1)}{\Pr(\phi(Z)=0)}$$

Let $\lambda = \Pr(\phi(Z)=1)$, $\pi = \Pr(\Theta=1)$, then

$$b = \frac{\pi}{\lambda(1-\lambda)}(\Pr(\phi(Z)=1|\Theta=1)-\lambda)$$

and noting that λ can be written as

$$\lambda = \Pr(\phi(Z)=1|\Theta=0)(1-\pi) + \Pr(\phi(Z)=1|\Theta=1)\pi$$

we have

$$b = \frac{\pi(1-\pi)}{\lambda(1-\lambda)}(1 - \psi_o - \psi_c)$$

where ψ_o, ψ_c are the omission and commission error rates of the classifier, viz,

$$\psi_o = \Pr(\phi(Z)=0|\Theta=1)$$
$$\psi_c = \Pr(\phi(Z)=1|\Theta=0)$$

Letting $\rho_1 = \text{var}(Z)/\lambda(1-\lambda)$ and $\rho_2 = \text{var}(Y)/\pi(1-\pi)$, we have

$$R^2 = (1 - \psi_o - \psi_c)^2 \frac{\pi(1-\pi)}{\lambda(1-\lambda)} \frac{\rho_1}{\rho_2}$$

Thus, the gain in sampling efficiency over using Y as the estimator depends upon the omission and commission error rates of the classifier, the probabilities π and λ, and two terms, ρ_1 and ρ_2 which are measures of the effect of clustering pixel observations into sampling units. The terms ρ_1 and ρ_2 are analogous to the intracluster correlation coefficients discussed in Cochran (2).

It is interesting to compare the sampling efficiency of the regression estimator E(Y) with that of a stratified estimator. Suppose the classifier is used to stratify the sampling frame into two strata; one stratum being the portion of the frame classified as the crop of interest and the other as the stratum of all other material. If instead of grouping the pixel observations into sampling units, they are randomly allocated in a pps (probability proportional to size) fashion to the strata, we could then take as an estimator of the proportion of acreage that is the crop interest

$$\hat{p} = \frac{n_{11}}{n_1}\lambda + \frac{n_{10}}{n_o}(1-\lambda)$$

Here n_1, n_o are the respective numbers of samples allocated to the stratum classified as the crop of interest and the other stratum, and n_{11} and n_{10} are the respective number of samples in n_1 that are in fact the crop of interest and the number of samples in n_o that are in fact other material. Here $n_1 = \lambda n$, $n_o = (1-\lambda)n$ where n is the sample size of ground enumerated pixels. For this case

$$\text{var}(\hat{p}) = R_1 \frac{\pi(1-\pi)}{n}$$

where

$$R_1 = 1 - (1 - \psi_o - \psi_c)^2 \frac{\pi(1-\pi)}{\lambda(1-\lambda)}$$

This expression is the same as the expression derived by Tenebein (3); however, Tenebein considers the case where the samples are thrown at random and not in a pps fashion. In his case R_1 has the above form in the limit (i.e., as $n \to \infty$).

7.0 CONCLUDING REMARKS: In this paper we have derived an expression for the regression function which indicates how certain properties of the classifier will affect the form of the regression function. Unfortunately, it is not always possible to achieve these properties with a given set of observations. For example, consider the regression function E(X|Y). We saw that a classifier which acts as an ancillary statistic on a family of crop class conditional distributions of the satellite measurements, whose parameters are functions of y (i.e., the family can be indexed on y which is the proportion of the crop of interest), makes E(X|Y) a linear function. This amounts to holding the **classifier omission** and commission error probabilities constant across B_y-groups of sampling units. Given an area, the satellite measurements of the crops in the area, along with the classifier design, determine these error probabilities. And, since one has no control over the measurements, it is not always possible to achieve the desired ancillary properties. Nevertheless, it may be possible to take certain steps which can help, such as stratifying the sampling frame on variables which can influence these conditional distributions. Strata which are homogeneous in crop mix, crop condition, and crop stage may be desirable (4).

If we want to keep E(Y|X) approximately linear, then our numerical example suggests that the Bayes classifier may be a better choice than the maximum likelihood classifier. But here again, the design of such a classifier depends in part upon variables whose values are not known in advance; e.g., the true proportions of the crop of interest in the sampling units. If the survey is repeated each year, then it may help to use previous year data over the sampling units to obtain a rough estimate of the desired crop proportions.

Finally, we remark that the expression we derived assumed that an infinite sample is available from each sampling unit. Thus, the treatment avoids certain estimation considerations which for small sample sizes may have dominant effects.

8.0 ACKNOWLEDGEMENT: The author wishes to thank Dr. Helen C. Takacs, Mississippi State University for providing the numerical example.

9.0 REFERENCES

1. Hanuschake, G.; Sigman, R.; Craig, M.; Ozga, M.; Luebbe, R.; Cook, P.; Klesveno, D.; Miller, C.: "Obtaining Timely Crop Area Estimates Using Ground-Gathered and Landsat Data," ESCS/USDA, Tech. Bulletin No. 1609, Aug. 1979.
2. Cochran, W.G.: Sampling Techniques. Third Ed., John Wiley and Sons.
3. Tenebein, A.: "A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications," JASA, Vol. 65, No. 33, Sept. 1970, pp. 1350-1361.
4. Hay, C.M.; Thomas, R.W.: "Application of Photointerpretive Techniques to Wheat Identification, Signature Extension, and Sampling Strategy," NASA Contract NAS9-14565, Space Sciences Laboratory, University of California, Berkeley, May 14, 1976.

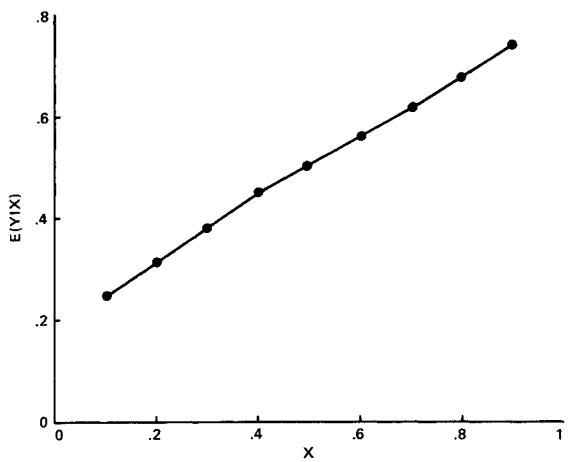

Figure 1(a).– E(Y|X) FOR THE BAYES CLASSIFIER.

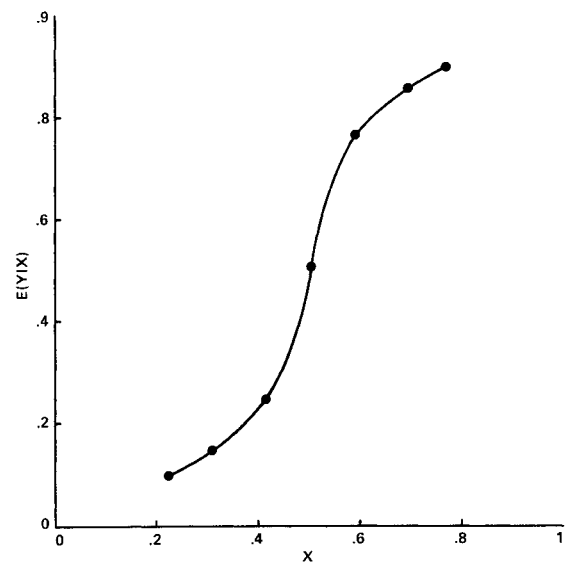

Figure 1(b).– E(Y|X) FOR THE MAXIMUM LIKELIHOOD CLASSIFIER.

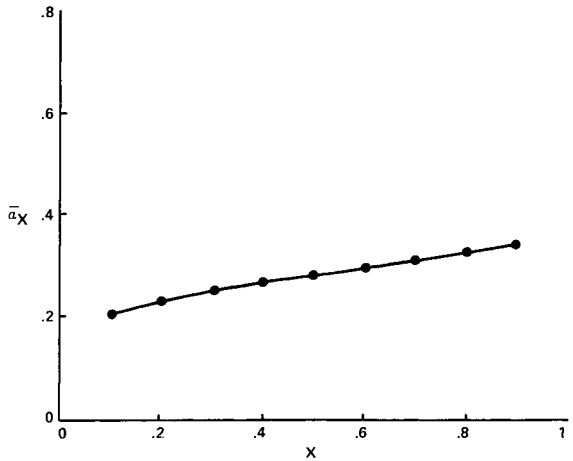

Figure 1(c).– $\bar{\alpha}_X$ FOR THE BAYES CLASSIFIER.

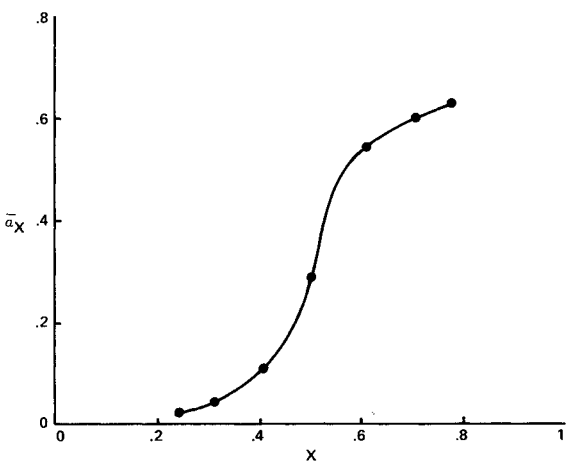

Figure 1(d).– $\bar{\alpha}_X$ FOR THE MAXIMUM LIKELIHOOD CLASSIFIER.

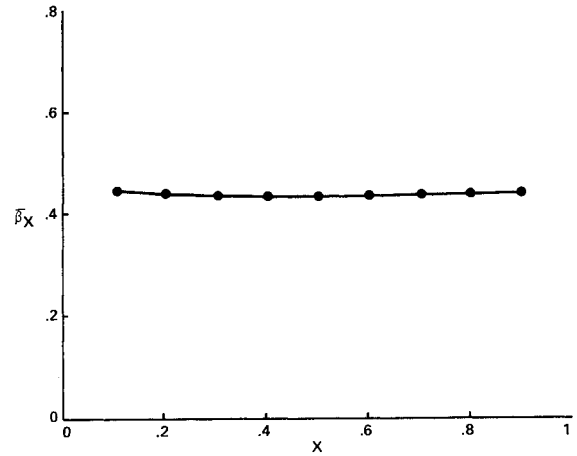

Figure 1(e).– $\bar{\beta}_X$ FOR THE BAYES CLASSIFIER

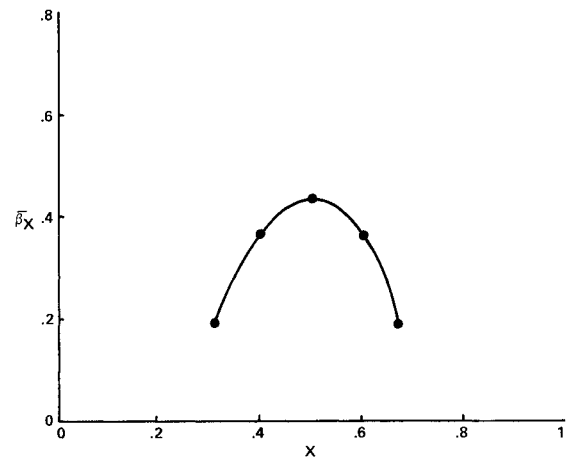

Figure 1(f).– $\bar{\beta}_X$ FOR THE MAXIMUM LIKELIHOOD CLASSIFIER