A NONLINEAR TWO-PHASE PREDICTOR FOR SOIL SURVEY UPDATES

Pamela J. Abbitt, F. Jay Breidt, and Sarah M. Nusser, Iowa State University
Pamela Abbitt, 208A Snedecor, Iowa State University, Ames, IA, 50011 pja@iastate.edu

Key Words: Tobit model, small area estimation

Abstract:

The National Cooperative Soil Survey is responsible for constructing soil maps detailing the location of soil series throughout the U.S. Methods for updating soil surveys have generally been based on purposive sampling. Recent developments in GIS and GPS technologies have made it possible to collect data at randomly located points. In a pilot project in western Iowa, a multi-phase sampling approach is being used to update soil surveys in two counties. A two-dimensional Markov chain sample is selected to provide a dense, geographically dispersed first phase sample. This sample is stratified according to soil map units which correspond to a specific soil series, slope class and erosion phase. A three-phase sample is drawn within each stratum from the first phase sample points. An objective of the soil survey is to summarize characteristics of soil map units using means, percentiles and other distributional parameters. We describe regression estimation procedures for estimating parameters that characterize map units. Special features of the data include truncated distributions for which nonlinear predictors are appropriate.

1. Introduction

The National Cooperative Soil Survey (NCSS) is a cooperative program involving the USDA and a state agency, often the state's Agricultural Experiment Station. The NCSS program is charged with constructing soil maps detailing the location of soil series thoughout the U.S. For each county, reports which contain soil maps and descriptions of each soil map unit within the county are generated. These maps are periodically updated through the NCSS program to provide current information on characteristics for different soils. Updates are based on soil surveys involving extensive field work. Traditionally, data on the distribution of soil properties is gathered during the survey using purposive sampling methods. This information is used by contractors, farmers and others for land use planning purposes and by scientists to develop models based on soil characteristics.

The county reports contain maps of soil polygons. The boundaries of these polygons are determined through soil surveys. Each soil polygon is called a *delineation*. During the survey, each delineation is assigned a *map unit symbol* (MUS) which is a 3-5 character alphanumeric code that identifies the dominant *soil series* (the name of the soil), *slope class* (the steepness of the slope), *erosion phase* (the state of erosion of the delineation) and *inclusions* (contrasting soils) within the delineation.

County reports also contain soil map unit descriptions. A *soil map unit* (SMU) consists of all delineations with the same MUS. These descriptions include a representative profile of the soil as well as location and geological history of the soil.

2. MLRA 107 Update

A stratified multi-phase sampling plan was used to conduct soil survey updates (Abbitt and Nusser, 1995). The universe for an update is the land surface of a county, although the proposed design may be used with any well-defined region, such as a Major Land Resource Area (MLRA). Sample sizes were allocated to each SMU approximately proportional to the surface area of the SMU.

To implement the design, an overly dense set of points is selected throughout the region. The selection procedure is a Markov Chain (MC) design for one-per-stratum sampling as described in Breidt (1995). This dense set of points is the phase 0 sample. For this phase, the SMU from the previous soil survey is obtained using digitized maps. In subsequent phases, subsamples stratified by SMU are drawn from the MC sample.

The design consists of three phases: surface horizon points, full profile points, and lab points. In all phases, values of data collection variables are recorded for each horizon. A *horizon* is a layer of soil which differs from the adjacent layers in physical, biological or chemical properties. In the first phase, information is collected on the physical characteristics of the point that are easily determined from the *surface horizon*. The surface horizon consists of the uppermost horizon at the point. In the second phase, field-observable data is collected on all horizons up to a depth of 48 inches where possible. In the third phase, lab determinations are made on soil samples taken from the field.

3. Small Area Estimation

There are often large numbers of SMUs in a survey area. For example, in Crawford County, Iowa, there are approximately 80 SMUs with a total sample size of approximately 2100 points. Practical constraints on personnel, time and money prevent us from choosing a design with adequate sample sizes in each of these domains. This suggests that small area estimation techniques may be useful for estimation at the SMU level.

The structure of soil taxonomy suggests useful groupings for modeling. Soils are classified using a hierarchical structure. This hierarchy is seen in the notation used for the MUS in Iowa. The numbers at the beginning of the MUS identify the dominant soil series. The letter indicates the slope class of the delineation. If no letter is present, the slope class is assumed to be A (0-2% slope). A number at the end of the MUS gives the erosion phase of the delineation. The erosion is assumed to be none to slight if there is no number. For example, the MUS 9D2 is in the Marshall series on a D slope (9-14%) with moderate erosion. Combinations of slope class and erosion phase are called phases of a soil series. We will use all phases of the Marshall soil series in the estimation example which follows.

4. Depth to first B horizon

The number, depth, and names of horizons change from point to point in the same SMU. Horizon names indicate some characteristics of the horizon, such as organic matter content and textural composition. The names follow roughly an A,B,C ordering (A horizons near the surface, followed by B horizons, followed by C horizons). There are many variations on A, B, and C horizons (e.g. Bt, AB) and other letters are also used to name horizons (e.g. O, E). The depth to first B horizon, *bdepth*, is of interest to soil scientists because it is related to depth to maximum clay content. The variable *bdepth* is recorded for phase II points.

The multi-phase design and the relationships among variables recorded in different phases suggest using regression estimation. The depth of the surface horizon, *sdepth*, provides a lower bound on *bdepth*. Since *sdepth* is not constant, we will model $d = bdepth - sdepth$. Approximately 25% of the observations have $d = 0$, with the rest of the values ranging from 3 to 44 inches.

The difference, d, is non-negative. Using a linear regression model may result in negative predictions for some values of d for phase I points. To avoid any negative predictions and to address the censoring problem, we will use a Tobit model.

5. Tobit models

Tobit models are also referred to as models with limited dependent variables, because we observe only a censored version of the dependent variable. In a basic Tobit model, we assume that there is an underlying linear model, but we observe non-negative values of the dependent variable. The underlying model is

$$d^* = x'\beta + \epsilon.$$

We observe

$$d = \begin{cases} d^* & \text{if } d^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Assuming $\epsilon \sim N(0, \sigma^2)$, we can obtain a maximum likelihood estimate of β. This basic model can be extended to allow for heteroskedastic models.

Under the Tobit model,

$$\mathrm{E}(d|d > 0) = x'\beta + \sigma \cdot f/F \qquad (1)$$

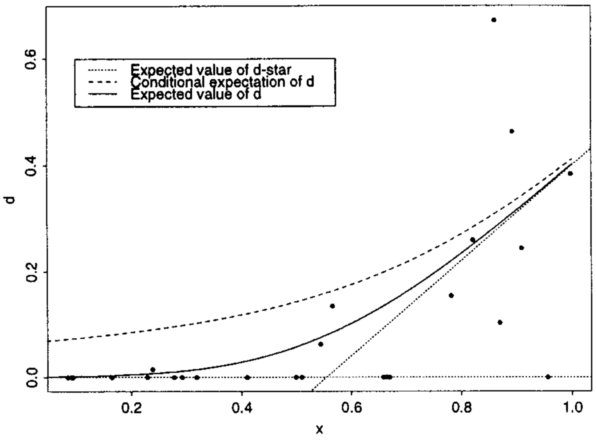
$$\mathrm{E}(d) = F \cdot \mathrm{E}(d|d > 0) \qquad (2)$$

where f and F are the probability density function and cumulative distribution function, respectively, of a standard normal distribution evaluated at $x'\beta/\sigma$ (Judge et al, 1988). Equation 2 is used to predict values of d, after obtaining estimates of β and σ. Figure 1 illustrates these expectations. Predicted values of d are added to *sdepth* to obtain a predicted value for *bdepth*, $\hat{b}$.

6. Estimators

We will compare two estimators of the mean of b for SMU i: the direct estimator, $\bar{b}_i$, and the regression estimator, $\tilde{b}_i$. Recall that design strata are SMUs from the previous soil survey. The strata will be indexed by h. We will let s_h and r_h denote phase I and II samples, respectively, for stratum h. The design is such that sampling weights are constant within stratum. During the survey, the new MUS is recorded for each point. These new SMUs are the

Figure 1: Expectations associated with the Tobit model, and simulated values of d



groups for which we want estimates. New SMUs are indexed by i.

The direct estimator, $\bar{b}_i$, is a combined ratio estimator computed for each SMU using only the observations which fall in the SMU. Only phase II points are used in this estimator, since these are the only points for which we have observations on *bdepth*. The estimator is

$$\bar{b}_i = \frac{\sum_{h=1}^{H} \sum_{j \epsilon r_h} w_{IIh} b_{hj} I_{ij}}{\sum_{h=1}^{H} \sum_{j \epsilon r_h} w_{IIh} I_{ij}},$$

where w_{IIh} is the phase II sampling weight for points in stratum h, b_{hj} is the value of *bdepth* for point j in stratum h, and I_{ij} is one if j falls in new SMU i and zero otherwise.

The regression estimator, $\tilde{b}_i$, is a composite estimator. Phase II points are used to fit a Tobit model. This model is then used to predict values of *bdepth* for all phase I points. The resulting estimator is

$$\tilde{b}_i = \frac{\sum_{h=1}^{H} \left\{ \sum_{j \epsilon s_h} w_{Ih} \hat{b}_{hj} I_{ij} + \sum_{j \epsilon r_h} w_{IIh} e_{hj} I_{ij} \right\}}{\sum_{h=1}^{H} \sum_{j \epsilon s_h} w_{Ih} I_{ij}},$$

where w_{Ih} and w_{IIh} are the phase I and phase II sampling weights, respectively, and $\hat{b}_{hj}$ and e_{hj} are the predicted value and residual, respectively, from a fitted model for point j in stratum h.

Variance estimators are calculated using standard Taylor linearization techniques for stratified ratio and regression estimators. In order to obtain variance estimates for those SMUs with only one phase II point, points were pooled within slope classes. For a particular slope class, variance estimates are needed for each old SMU (design stratum). Some of these estimates cannot be calculated due to sample size. For those strata, the mean of the variance estimates from other strata is used as an estimate.

7. Example

Data collection is being conducted in Crawford County, Iowa. The total sample size is 2131 points. The current data set is partial and not necessarily representative. However, Marshall soils were targeted for data collection for the purposes of developing estimation techniques. Data has been collected at 146 points which have Marshall soils. Preliminary sampling weights were calculated using old SMU acreage as auxiliary information.

Table 1: Estimated mean of *bdepth* for soil map units in the Marshall series. Estimated standard errors are given in parentheses, n_{s_h} and n_{r_h} are the phase I and II sample sizes, $\bar{b}_i$ is the direct estimator, $\tilde{b}_{1i}$ and $\tilde{b}_{2i}$ are regression estimators based on two different Tobit models, and $\tilde{b}_{3i}$ is a regression estimator based on a linear model.

SMU	n_{s_h}	n_{r_h}	$\bar{b}_i$	$\tilde{b}_{1i}$	$\tilde{b}_{2i}$	$\tilde{b}_{3i}$
9B	19	4	17.0 (1.6)	17.8 (1.4)	17.8 (1.5)	17.8 (2.2)
9B2	4	1	10.0 (6.3)	11.6 (6.5)	11.9 (7.1)	12.2 (10)
9C	5	1	13.0 (18)	15.4 (18)	15.1 (18)	15.0 (18)
9C2	57	15	12.0 (3.0)	12.4 (3.0)	12.3 (2.9)	12.3 (2.9)
9D	3	0	·	25.4	20.0	19.8
9D2	50	12	18.5 (7.0)	18.8 (6.7)	16.6 (4.2)	17.2 (4.3)
9D3	2	1	17.0 (28)	18.1 (56)	12.5 (32)	10.4 (33)
9E2	4	1	12.0 (32)	11.4 (45)	7.8 (28)	-6.7 (29)
9E3	2	0	·	14.0	6.0	2.0

Table 1 shows sample sizes and estimates. Some of the standard errors could not be calculated with the method described above. Those SMUs without estimated standard errors do not have any phase II points. The two different Tobit models use different covariates. The first model includes only class variables for slope class and erosion phase. The second model includes these variables and a class variable for horizontal slope shape. The third regression estimator, $\tilde{b}_{3i}$, is based on a linear predictor using the same covariates as the second Tobit model.

The regression estimators based on these models allow us to obtain estimates for SMUs which have no phase II points. However, care must be used in constructing these models. For example, in SMU 9E2, $\tilde{b}_{3i}$ is negative. This is one disadvantage of using a linear predictor. The variance estimates are not much lower, if at all, for the regression estimators in those SMUs for which we can calculate direct estimates. This may be due to lack of fit. However, providing non-negative estimates for each SMU is an important advantage of the Tobit-based regression estimators.

The Tobit models were fit under the assumption of normality. Preliminary diagnostic plots indicate that this may not be appropriate. If the normality assumption is violated in the Tobit model, the estimate of β is inconsistent. Estimation methods which do not rely on this assumption may provide a better fit and result in lower variances for the regression estimator.

8. Summary

This procedure is part of a larger project to broaden the utility of soil surveys. Regression estimation and small area estimation are two techniques that are appropriate in this setting. Linear models are not appropriate for describing relationships between many variables of interest in this project. Using nonlinear models in conjunction with regression estimation and small area estimation may lead to improved estimates.

9. References

Abbitt, P.J. and S.M. Nusser. (1995). Sampling approaches for soil survey updates. *ASA Proceedings of Statistics and the Environment Section*, 87-91.

Breidt, F.J. (1995). Markov chain designs for one-per-stratum spatial sampling. *ASA Proceedings of Survey Research Methods Section*, vol I, 356-361.

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lütkepohl, and T.C. Lee. (1988). *Introduction to the Theory and Practice of Econometrics.* Wiley, New York.