Modeling Which Farms Are Not Covered by a Census List Using an Area-Frame Survey

Ted Chang
University of Virginia

Phillip S. Kott
National Agricultural Statistical Service

Abstract

We summarize the results of efforts to model the probability of a farm not being on the Census Mailing List maintained by the National Agricultural Statistics Service (NASS). The analyses were based on data from the NASS area-frame sample and treated that frame as complete. The predictive covariates considered involved total sales, type of farm, acreage, operator characteristics (gender, Hispanic status, race, and whether the principal occupation of the principal operator is farming), number (if any) of equine on the farm, and the Area-Frame-Survey stratum.

1. Introduction

The National Agricultural Statistics Service (NASS) based the 2002 Census of Agriculture on a list of farms called the Census Mailing List. Some 18% of all farms were *not on the List* or *NML*. We summarize research aimed at modeling the causes of farms not being on the Census Mailing List. Chang and Kott (2004) describes this effort in greater detail. NASS is using the results of this research to sharpen its list-building efforts for the 2007 Census of Agriculture and to more efficiently allocate the 2007 area sample which will be used to measure the undercoverage of the 2007 Census.

Our underlying model stipulates that each farm k in the US has a Poisson (i.e., independent) probability P_k of being NML. This probability depends on various covariates, $X_{k1}, ..., X_{kQ}$, associated with the farm:

$$P_k = f(X_{k1}, ..., X_{kQ}), \tag{1}$$

A research goal was the development a procedure for choosing good covariates and an appropriate function f(.).

The data used for this analysis came from the June 2002 Area Frame Survey and its fall supplement, the Agricultural Coverage Evaluation Survey (ACES). We will refer to this tandem as the AFS in what follows. We assume every farm located in the 48 contiguous states in 2002 had a positive, calculable probability of selection in the AFS.

Section 2 describes a logistic model for a farm being NML in California. A modification of the stepwise selection algorithm for model selection was used in model fitting with modifications incorporating design-based variance estimates. Section 3 explores of appropriateness of using estimates from logistic modeling as a weighting tool, while Section 4 investigates alternative link functions, namely, the probit, log-log, and complementary log-log. Section 5 describes fitting a model to the entire contiguous US conducted after Chang and Kott (2004). Section 6 provides a discussion of that fitting and considers future areas of research.

The Logistic Regression Model

For a farm with covariates $X_{k1}, ..., X_{kQ}$, the logistic NML model is

$$\log[P_k/(1 - P_k)] = \beta_0 + \beta_1 X_{K1} + ... + \beta X_{kQ}, \tag{2}$$

where the $\{\beta_q\}$ are unknown constants estimated when the model in equation (1) is fit with a logistic regression routine. If b_q is consistent estimator for β_q (q = 1, ..., Q), then

$$p_k = \exp\{b_0 + b_1 X_{K1} + ... + bX_{kQ}\}/[[1 + \exp\{b_0 + b_1 X_{K1} + ... + bX_{kQ}\}] \tag{3}$$

is a consistent estimator for P_k.

We fit the California AFS data using design-based logistic regression (Binder 1983). The fitted "no-strata" model was

$$\begin{aligned}\log[p_k/(1-p_k)] = {} & 2.442 - 1.035\,\texttt{sales5K} - 0.813\,\texttt{sales50K} - 1.788\,\texttt{sales1000K} \\ & - 1.251\,\text{CHRS} - 0.909\,\text{CCENFRUT} - 2.618\,\text{CCENCOTT} + 0.966\,\text{CCENSHEP} \\ & + 2.104\,\text{CCENAQUA} - 0.0362\,\texttt{age} + 1.140\,\texttt{hisp} + 1.028\,\texttt{asian} - 0.571\,\texttt{ocup},\end{aligned} \tag{4}$$

where sales5K = 1 when farm 2002 sales were at least \$5,000, 0 otherwise (sales50K and sales1000K defined conformally); CCENCHRS, CCENFRUT, CCENCOTT, CCENSHEP, CCENAQUA = 1 when the farm answered `YES' to corresponding survey question: that the farm produces Christmas trees, fruit and nuts, cotton, sheep, and products of aquaculture, respectively, and 0; CHRS = CCENCHRS - ftypCHRS, where ftypCHRS=1 when the farm listed Christmas trees as the primary source of sales, and 0 otherwise; age = age of principal operator rounded to a multiple of 10 years; hisp = 1 when principal operator has Hispanic background, and 0 otherwise; asian = 1 when the race of the principal operator was Asian, and 0 otherwise; ocup = 1 when the principal occupation of the principal operator was farming or ranching, and 0 otherwise.

3. An Experiment in Coverage Correction

An experiment to assess the accuracy of such correction was performed as follows. We let U be the farms in the California AFS and L be the subsample of U on the Census Mailing List The fitted model in equation (4) was used to generate p-values for each $k \in U$. Using the sampling weight w_k attached to each farm $k \in U$, the population total for various x-variables could be computed as $T_x = \sum_U w_k x_k$ and compared to the "estimated" value based only on farms in L: $t_x = \sum_L w_k x_k/(1-p_k)$.

Representative selected results, out of the 69 variables considered, are listed in Table 1. To facilitate comparison between t_x and T_x, the standard errors of t_x under the Poisson model (i.e., $\text{Var}(t_x) = \sum_U (w_k x_k)^2 p_k/(1-p_k)$), as well as the resulting t-ratios are given. Of the 69 variables considered, the worst result (as measured by the t-ratio) was for the variable strat11 (the most intensely agricultural area stratum), which is among the 13 shown on Table 1.

The conclusion of this experiment is that Poisson-probability-of-being-NML model and the fitted probabilities in equation (4) seem to fit the California AFS data.

4. Exploring Alternative Link Functions

Let us write the summation $\beta_0 + \beta_1 X_{K1} + \ldots + \beta X_{kQ}$ as η_k, so that the logistic model in equation (2) can be re-expressed as

$$P_k = \exp(\eta_k)/[1 + \exp(\eta_k)]. \tag{5.1}$$

The right hand side of equation (5.1) is called the link function.

We considered three other popular link functions. For comparison purposes, they have been normalized so P=0.5 and $dP/d\eta = 0.25$ when $\eta=0$. These link functions are

the *probit*: $P_k = \Phi\{[(2\pi)^{1/2}/4]\eta_k\}$ (5.2)
the *log-log*: $P_k = \exp\{-\log(2)\exp[(-\log(2))^{-1}\eta_k]\}$ (5.3)
the *complementary log-log*: $P_k = 1 - \exp\{-\log(2)\exp[(-\log(2))^{-1}\eta_k]\}$. (5.4)

All four link functions are monotonically increasing and S-shaped, approaching 0 as $\eta \to -\infty$ and 1 as $\eta \to \infty$. They differ primarily in the tails. The logistic and probit links are symmetric, the log-log link dies much quicker for large negative values of η than it does for large positive values and the complementary log-log link dies quicker for large positive values of η than it does for large negative values.

The coverage experiment described before was repeated for the link functions of equations (5.1) – (5.4). Representative results are shown in Table 2. For each link function, new coefficients were fit using the same variables used for the logistic-link fit. There does not appear to be substantial differences among the performance of the four links. We suspect this is because few farms have probabilities in the extreme tails.

Chang and Kott (2004) also describe model fitting with parameter estimates truncated to avoid overly large p_k values. The results are similar to those in Table 2. Since logistic model parameters are easier to interpret and computationally simpler to estimate, we fit logistic model exclusively for the remainder of the research.

5. Fitting the Logistic Model at the US Level

Chang and Kott (2004) discuss in some detail fitting data from the union of three Midwestern states and then the 48 contiguous states as a whole. A later US fit to be used by NASS in the area-sample allocation program is described below.

The variables were first organized into these groups: sales variables, farm-type variables, land variables, variables related to equine operations, operator- characteristic variables, state, and strata. The approaches used to fit the main effects and the interaction terms were slightly different. The basic approach was a modification of the stepwise regression algorithm, starting with a model of intercept only.

For the main effects, the procedure was iterative in the groups. At each iteration, the groups were ordered by their significance level in a design-based Wald test comparing a full model consisting of the current model plus the variables in the group and a reduced model of the current model alone. The most significant group was chosen. At this point a stepwise regression procedure was used to choose variables within the group to add to the current model. These tests were all conducted at a .05 significance level.

Starting with a model of intercept only, the most significant predictor group was sales. Of the 10 possible sales variables, 6 were found to be sufficient to account for the predictive power of the group.

In the end, all groups were significant. They entered in this order: sales, land variables, operator characteristics, variables related to the type of equine operation, farm-type variables, state, and strata. The algorithm was rerun to check for variables that might become either significant or insignificant. In this way, 38 out of 129 possible main effects entered into the model.

We found that the inclusion of too many interaction terms lead to models with numerical instability and poor predictive power. Thus, only two-way interaction terms including a state variable as one of the components were considered. In addition, a very conservative algorithm was used to fit them. Two types of standard errors were computed, a design-based standard error and a model-based standard error (where clustering and stratification were ignored and sampling weights treated as nuisances). Generally speaking, the model-based standard error was larger than the design-based standard error. The addition in the model of a term with a large ratio of model-based to design-based standard error lead to numerical instability and poor predictive power.

The groups were listed in the same order that their main effects entered the model. For each group, 43 models, each consisting of the interactions of one of the 43 states with the variables in the group, were considered for addition into the current model (NASS treats New England as a single state and has no AFS in Alaska and Hawaii.)

Of these 43 models, the ones with a (model-based) significance level below .001 (this is a Bonferroni correction with an overall significance level of .043) were considered as possible interaction terms. Backwards elimination, with a significance level of .05, was used to winnow down the interaction terms of this type. Interaction terms, in which the model-based standard error exceeded the design based standard error by a factor of 2 or more, were also eliminated from consideration.

This method added only five of 3,698 possible interaction terms; however, it was later found that even this, extremely conservative procedure, produced two probable erroneous terms, which we discuss later. It should be noted that the procedure employed in July 2004 was based upon a different standard error and produced significantly more interaction terms.

At this point, the model included the variables female, Hispanic, and black. Mathematically, this forces the nationwide estimate of NML farms with principal operators in these population groups to be the same as the crude estimator obtained by adding the weights of the area-frame-survey NML farms. To maintain this consisting for Asians, `asian` was added to the model by fiat. Observe this parameter does have a relatively large estimate in absolute terms.

Although the nationwide estimates of the model and the crude estimators will agree for population groups like blacks and Asians, estimates for lower level units (e.g. blacks in Texas) will not agree; the model in effect uses other variables, such as sales and land size, to smooth out the estimates for these population groups among at lower levels of aggregation.

Finally we found that the interaction terms for the states of CA and AR with the variable for participation in Conservation or Wetland Reserve Programs (clancrp2) caused inconsistent predictions of NML status for farms in these states that participate in the reserve programs. We believe that this is due to the area-frame survey having only two NML farms in the reserve programs in CA and five in AR. These farms had relative large weights due to their small size. For this reason the two state/CRP interaction terms were deleted from the model.

6. Discussion

Many of the results in Table 3 are not surprising. The larger the annual sales, the more likely a farm is on the Census Mailing List. Farms with older operators and with operators who consider themselves primarily farmers are more likely to be on the List.

One issue of some debate before this analysis was whether the well-known tendency for the List to be missing black operators simply reflected the types of farms blacks operated. That proves *not* to be the case. In fact, all other things being equal, having a black operator turned out to be one the best predictors of a farm being NML.

We also learned that the 2002 area-sample allocation favored precisely those farms very likely to be on the List – farms in in intensely agricultural area strata.

Chang and Kott (2004) discuss at some length what they call the "hidden-small-cell problem" and the related issue of the numerical instability of some parameter estimates. That these problem arose in such a large data set (46,000 observations) with a moderate number of variables (around 50) was surprising. A deeper exploration into the asymptotic failings of the design-based methods they primarily used would be helpful as well as investigations into model-based and model-assisted alternatives.

Small farms in the AFS data set often have relatively large sampling weights. This is because the AFS is actually a sample of area segments. The portion of a farm within a selected segment is called a "tract." A farm with a tract within a sampled area segment is given a sampling weight equal to the product of, 1, the inverse selection probability of the area segment, and, 2, the ratio of the area in the tract to the area in the whole farm. Often, for small farms, this ratio is close to 1, but for large farms, it can be quite small (less than .01). Thus, it is possible that for some parameters small farms are excessively influential. The need for techniques to uncover such situations is compelling.

Acknowledgements

The authors wish to thank Herb Eldridge of NASS for preparing the data file and Bill Wigton, also of NASS, for his many insights into the Area Frame Survey.

References

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292

Chang, T. and Kott, P.S. (2004). *Modeling NML Using the Area Frame Survey.* http://www.nass.usda.gov/research/reports/nml0826.pdf

Folsom, R. E. and A. C. Singh (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *ASA Proceedings of the Survey Research Methods Section*, 598-602.

Garren, S. and T. Chang (2002). Improved ratio estimation for telephone surveys adjusting for noncoverage. *Survey Methodology*, **28**, 63-76.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall.

Sarndal, C.-E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling.* Springer-Verlag.

Silvey, S. D. (1970). *Statistical Inference*, Chapman and Hall.

Table 1: List-based Estimates

Variable	T_x	t_x	$\sqrt{\mathrm{Var}(t_x)}$	t-ratio	variable description
1	68896.	68727.	2991.	-0.056	number of farms
CCENGRAN *	5701.3	5882.3	371.8	0.487	farms which grow grains
CCENCHRS	299.20	190.17	198.18	-0.550	grows Christmas trees
CCENCATL *	14378.	14461.	1284.	0.065	raises cattle
ftypFRUT	29192.	29267.	974.	0.077	primary sales are fruit and nuts
sales10K *	39564.	38371.	1087.	-1.097	annual sales at least $10K
sales100K *	19912.	19576.	574.	-0.585	annual sales at least $100K
sales1000K	5860.6	5884.8	106.8	0.226	annual sales at least $1,000K
hisp	8088.8	7844.0	1568.3	-0.156	operator is Hispanic
ocup	37002.	38343.	1405.	0.954	operator’s principal occupation is farming
LEQUIOWN *	87805.	78990.	10250.	-0.860	number of horses owned
CLANDTOT *	22962.	23375.	794.	0.521	total land area (in units of 1,000 acres)
strat11 *	17268.	19540.	1030.	2.207	number of farms in area stratum 11 (most intensely agricultural)

* Terms marked with an asterisk (*) do not appear in the fitted model

Table 2: Alternative Link Functions

Variable	T_x	Logistic	Probit	Log-log	Complementary log-log
1	68896.	68727.	68666.	68397.	68754.
CCENGRAN	5701.3	5882.3	5887.0	5951.6	5853.9
CCENCHRS	299.20	190.17	191.92	201.09	185.01
CCENCAT	14378.	14461.	14368.	14388.	14393.
ftypFRUT	29192.	29267.	29315.	29450.	29172.
sales10K	39564.	38371.	38437.	38710.	38447.
sales100K	19912.	19576.	19624.	19803.	19559.
sales1000K	5860.6	5884.8	5918.9	5960.3	5869.1
hisp	8088.8	7844.0	7824.3	7537.6	7835.0
ocup	37002.	38343.	38449.	38908.	37697.
LEQUIOWN	87805.	78990.	78462.	77910.	78790.
CLANDTOT	22962.	23375.	23426.	23722.	23390.
strat11	17268.	19540.	19490.	19105.	20126.

Table 3: The US-Level Estimated Logistic Model Parameters

intercept	1.645	1 for all data points
sales 1K	-0.663	1 if sales greater or equal to 1K; 0 otherwise
sales2.5K	-0.546	1 if sales greater or equal to 2.5K; 0 otherwise
sales10K	-0.452	1 if sales greater or equal to 10K; 0 otherwise
sales50K	-0.302	1 if sales greater or equal to 50K.5; 0 otherwise
sales250K	-0.598	1 if sales greater or equal to 250K; 0 otherwise
age	-0.203	Age in decades (2 = under 25; 3 = 25-34; etc.)
ocup	-0.229	1 if principle occupation was farming/ranching; 0 otherwise
female	0.313	1 if principle operator was female; 0 otherwise
hisp	0.386	1 if principle operator was Hispanic; 0 otherwise
black	0.916	1 if principle operator was at least part Black; 0 otherwise
asian	0.617	1 if principle operator was at least part Asian; 0 otherwise
leqoper4	0.275	1 if operation had equine for personal use; 0 otherwise
leqoper5	0.512	1 if operation had equine for other reasons; 0 otherwise Other than being a farm/ranch, a breeding service, a boarding, training or riding facility, or a place to keep equine for person use.
CCENGRAN	-0.225	1 if operation had grain crops; 0 otherwise
CCENOTHC	-0.296	1 if operation had other crops or hay; 0 otherwise
CCENCATL	-0.246	1 if operation had cattle; 0 otherwise
ftypGRAN	-0.364	1 if a grain-crops operation; 0 otherwise
ftypTOBA	-1.093	1 if a tobacco operation; 0 otherwise
ftypCATL	-0.303	1 if a cattle operation; 0 otherwise
ftypSHEP	-0.371	1 if a sheep operation; 0 otherwise
ftypFRUT	-0.466	1 if a fruit operation; 0 otherwise
ftypHOGS	-0.406	1 if a hog operation; 0 otherwise
ftypMILK	-0.622	1 if a milk operation; 0 otherwise
HOGS	0.400	1 if operation had hogs but not a hog operation; 0 otherwise
MO	0.483	1 if farm was in Missouri; 0 otherwise
TX	0.278	1 if farm was in Texas; 0 otherwise
NE	-0.475	1 if farm was in Nebraska; 0 otherwise
MS	0.388	1 if farm was in Mississippi; 0 otherwise
N_Eng	0.392	1 if farm was in New England; 0 otherwise
FL	0.313	1 if farm was in Florida; 0 otherwise
strat10s	-0.200	1 if operation was in stratum 11-19 (intensely agricultural);0 otherwise
clancrp2	-1.379	1 if operation had CRP land; 0 otherwise
clantot2	-0.294	(Total land acres)/1000, but no greater than 10
croplan2	-0.631	(Cropland acres)/1000, but no greater than 10
clantot3	0.0344	(clantot2 - 2)^2
croplan3	0.0917	(croplan2 - 1)^2
CA:clantot2	0.206	CA × clantot2
CO:clantot2	0.2 26	CO × clantot2