Modeling NML Using the Area Frame Survey

Theodore Chang *
University of Virginia
National Agricultural Statistical Service
tcc8v@virginia.edu

Phillip S. Kott
National Agricultural Statistical Service
Phil_Kott@nass.usda.gov

August 27, 2004

Abstract

This report outlines the results of various experiments to model the probability of an area-frame farm *not* being on the census mailing list (NML) using covariates such as total sales, type of farm, acreage, various operator characteristics (gender, Hispanic status, race, and whether the primaryl occupation of the principal operator is farming), number (if any) of equine on the farm, and, optionally, Area-Frame-Survey stratum.

Three sets of experiments were conducted. The first used California data only. The second used data from three states - Illinois, Indiana, and Iowa, while the third used data from the entire 48 contiguous states (there is no Area Frame Survey in either Alaska or Hawaii).

The statistical methodology employed was logistic regression with a modification of stepwise regression for variable selection. Standard errors were estimated using design-based linearization methods.

Conclusions are drawn about the nature of 2002 NML farms. Some are as expected (farms with small total sales are more likely to be NML). Others are surprising (holding all other factors constant, point farms are not sgnificantly more likely to be NML than other farms with less than $2,500 in annual sales).

A number of methodological problems needing further research are suggested as a result of this study.

*The empirical investigations described in this report where conducted by Professor Chang while he was an ASA research fellow at NASS from January to July of 2004. Although much was accomplished, there are limitations in the analysis due to the compact time frame.

1 Introduction and Summary

This study starts with a model stipulating that each farm has a probability p of *not* being on the census mailing list (i.e., of being NML - not on the mail list), and that this probability depends on various covariates $X_{+1}, \cdots, X_{+P}$ of the farm.[1] In other words, we assume a model of the form

$$p = f(X_{+1}, \cdots, X_{+P}).$$

The purpose of this study is to develop a procedure for choosing good covariates $X_{+1}, \cdots, X_{+P}$ and an appropriate function f.

The data used for this study 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.

In consultation with Herb Eldridge, candidate variables for the X_{+j} were chosen. These consisted of sales variables, farm-type variables, variables related to operator characteristics, variables related to equine ownership, AFS stratum, and variables related to total acreage.

One might suppose that the optimal strategy would be to include all possible variables. To see why this is not the case, consider the model fitted to the California data using only sales covariates. That model is[2]

$$\begin{aligned} \log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) &= 0.134 + 0.520 \cdot \text{sales1K} - 0.831 \cdot \text{sales2.5K} + 0.413 \cdot \text{sales10K} \\ &- 0.868 \cdot \text{sales5K} - 0.459 \cdot \text{sales25K} - 1.076 \cdot \text{sales50K} \\ &+ 0.584 \cdot \text{sales100K} - 0.741 \cdot \text{sales250K} + 0.711 \cdot \text{sales500K} \\ &- 1.816 \cdot \text{sales1000K}, \end{aligned} \tag{1.1}$$

where, for example, *sales10K* $= 1$ when the sales, as indicated by MFARMDEF[3], is *at least* \$10,000, and *sales10K* $= 0$ otherwise. For the model in equation (1.1), if a farm has sales of \$5,000, then

$$\begin{aligned} \text{sales1K} &= \text{sales2.5K} = \text{sales5K} = 1, \\ \text{sales10K} &= \text{sales50K} = \cdots = \text{sales1000K} = 0, \\ \text{and } \widehat{p} &= (1 + \exp[-(0.1337 + 0.5197 - 0.8312 - 0.8680)])^{-1} = 0.26, \end{aligned}$$

whereas if a farm has sales of \$10,000, then

$$\begin{aligned} \text{sales1K} &= \text{sales2.5K} = \text{sales5K} = \text{sales10K} = 1, \\ \text{sales50K} &= \cdots = \text{sales1000K} = 0, \\ \text{and } \widehat{p} &= (1 + \exp[-(0.1337 + 0.5197 - 0.8312 - 0.8680 + 0.4126)])^{-1} = 0.35. \end{aligned}$$

Now, the probability of NML status should decrease with increasing sales. It follows that the coefficients of all the sales variables should be negative. That is not the case in equation

[1] The notation we shall follow is that if we are referring to a specific farm i, then the probability is written p_i and the covariates $X_{i1}, \cdots, X_{iP}$. The generic probability is denoted p and the generic covariates by $X_{+1}, \cdots, X_{+P}$.

[2] In conformity with standard notation, we will use p for a hypothesized probability and $\widehat{p}$ for an estimate of p. The variable p is unknown, whereas $\widehat{p}$ is calculated from a sample.

[3] We use the convention that if a variable name appears in all upper case letters, it represents a standard NASS AFS variable name. Otherwise it represents a derived variable. In general, when a variable name appears in the text, it will either contain two or more upper-case letters or be italicized.

(1.1). Apparently, the model has overfit the data, resulting in the unsatisfactory behavior for some of the signs in the equation.

By contrast, when only the terms *sales5K*, *sales 50K*, *sales1000K* are used (as in the models of Section 2)

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = 0.3487 - 1.2665 \cdot \text{sales5K} - 0.9263 \cdot \text{sales50K} - 1.5842 \cdot \text{sales1000K}, \quad (1.2)$$

and all the estimated coefficients of equation (1.2) are negative as expected.

For this study, a modification of the stepwise selection algorithm for model selection was chosen with modifications incorporating design-based variance estimation and the grouping and types of variables in the NASS AFS questionnaire.

Three populations were considered: California only, the three states Illinois, Indiana, and Iowa taken together, and the 48 contiguous states (there is no AFS in either Alaska or Hawaii).

In general, sales was the most important predictor of NML status. Stratum, certain operator characteristics, and certain farm types were likewise significant predictors. For the 48-state data set, with its large size, acreage and state were also significant, but of less importance. These three models are discussed in Sections 2, 3, and 4 respectively. The results in Section 4, since they are the most broadly applicable, are given particular attention. Section 5 discusses statistical methodology in greater depth, while Section 6 offers some concluding remarks.

1.1 The logistic regression model

Suppose the covariates for the i-th sample farm are $X_{i1}, \cdots, X_{iP}$. The logistic regression model is that a farm with these covariates has a probability p_i of being NML, where p_i has the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{iP}. \quad (1.3)$$

The $\beta_0, \beta_1, \cdots, \beta_P$ are unknown constants which are estimated when the model in equation (1.3) is fit.

For example, the California 'no strata' model is

$$\begin{aligned}\log\left(\frac{\widehat{p}_i}{1-\widehat{p}_i}\right) \;=\; & 2.442 - 1.035 \cdot \text{sales5K} - 0.813 \cdot \text{sales50K} - 1.788 \cdot \text{sales1000K} \\ & -1.251 \cdot \text{CHRS} - 0.909 \cdot \text{CCENFRUT} \\ & -2.618 \cdot \text{CCENCOTT} + 0.966 \cdot \text{CCENSHEP} + 2.104 \cdot \text{CCENAQUA} \\ & -0.0362 \cdot \text{age} + 1.140 \cdot \text{hisp} + 1.028 \cdot \text{asian} - 0.571 \cdot \text{ocup},\end{aligned} \quad (1.4)$$

where

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 otherwise.

CHRS = CCENCHRS - ftypCHRS, where ftypCHRS=1 when the farm listed Christmas trees as the primary source of sales, and 0 otherwise[4].

age = age of principal operator rounded to a multiple of 10 years (a recode of MDEMOAGE).

hisp = 1 when principal operator has Hispanic background, and 0 otherwise (a recode of MDEMOHISP).

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 (a recode of MDEMOCUP).

One advantage of the logistic regression model in equation (1.3) is the ready interpretation of the coefficients β_j. Suppose we have two farms, indexed by i_1 and i_2, with identical covariates for $j = 2, \cdots, P$ (that is $X_{i_1 j} = X_{i_2 j}$ for $j = 2, \cdots, P$). Then, substituting into equation (1.3),

$$\frac{\left(\frac{p_{i_1}}{1-p_{i_1}}\right)}{\left(\frac{p_{i_2}}{1-p_{i_2}}\right)} = \exp(\beta_1). \tag{1.5}$$

In standard terminology $\frac{p_i}{1-p_i}$ is the *odds* of a farm being NML. We can say that in the model in equation (1.4) the odds of being NML is estimated to decrease by 30.4% for each 10 years of operator age (since $\exp(-.0362 \cdot 10) = .696 = 1 - .304$). Notice this this statement is true no matter what the values of the other variables in equation (1.4) are, as long as they are held constant. This easy interpretation of the coefficients would not apply if other link functions are used (see Section 1.4) which explains, in part, the popularity of logistic regression modeling.

A second advantage of regression-style modeling is that the effects of the various variables X_{+j} can be separated from each other. For example, it is highly probable that agricultural operations in which the primary occupation of the principal operator is farming have higher sales, on average, than farming operations in which farming is a secondary occupation of the principal operator. For the California AFS, the weighted-sample proportion of operations that were NML is $\widehat{\phi}_1 = 0.1946$ when the primary occupation of the principal operator was farming and $\widehat{\phi}_0 = 0.4545$ otherwise. In this case, the odds ratio

$$\frac{\left(\frac{\widehat{\phi}_1}{1-\widehat{\phi}_1}\right)}{\left(\frac{\widehat{\phi}_0}{1-\widehat{\phi}_0}\right)} = 0.290, \tag{1.6}$$

whereas (from the estimated coefficient of *ocup* in equation (1.4))

$$\exp(-0.571) = 0.565. \tag{1.7}$$

The difference between equations (1.6) and (1.7) is that the latter attempts to estimate the effect of principal occupation of the principal operator *after accounting for the effects*

[4]Because CCENCHRS is based upon current planting and future sales and ftypCHRS on previous year sales, it is possible for CCENCHRS=0, ftypCHRS=1, and CHRS=-1. Usually, however, CHRS=1 when the farm produces Christmas trees, but Christmas trees are not the primary source of sales, and 0 otherwise.

of the other variables, such as sales, in the model. Since a null hypothesis of no effect on NML translates to an odds ratio of 1, equation (1.7) is saying that part of the dramatic decrease in NML shown by equation (1.6) derives from farms with principal operators whose principal occupation is farming being different, on average, from farms for which this is not true. In essence, equation (1.7) tries to balance these two groups with respect to the other variables before calculating the NML proportions.

As a second example, although the questions on equine ownership appear highly predictive of NML, the gross sales are even more predictive, and the predictive power of equine ownership in California becomes statistically insignificant as soon as sales are entered into the model.

Finally, one might be tempted to cross tabulate the sample by the various factors in the model and determine the proportion of each cell in the cross tabulation that is NML. Unfortunately, there are 11 sales categories, 16 CCEN ("has a commodity") type variables, and six age categories. Thus a complete cross tabulation would have $11 \cdot 2^{16} \cdot 6 \cdot 2^3 = 34,603,008$ cells. Since the California AFS has only 1,488 farms, such a comprehensive cross tabulation would not yield meaningful results.

In fact, if a *saturated* model were used in equation (1.3), that is, one including all possible interactions, the results would coincide with those of a complete cross tabulation. Thus, depending upon the amount of data, models intermediate to complete cross tabulation can be fit by including some interactions (see Section 4.2).

1.2 Possible uses for an NML model

Let $q_i = 1 - p_i$ be the probability that the i-th farm is on the list. Let z_i be the characteristic of interest of the i-th farm, measured for each $i \in \mathcal{L}$, where $\mathcal{L}$ is the census mail list (CML) population. Then

$$\widehat{t}_{z0} = \sum_{i \in \mathcal{L}} \frac{z_i}{q_i} \tag{1.8}$$

is an unbiased estimate of

$$t_z = \sum_{i \in \mathcal{U}} z_i \tag{1.9}$$

the total of z_i for all $i \in \mathcal{U}$, where $\mathcal{U}$ is the population of interest, presumably the population of all farms.

We are assuming here a model that the list frame $\mathcal{L}$ can be regarded as a Poisson sample from $\mathcal{U}$ with probabilities q_i (see Särndal *et al.* [6]). [5] Similar models have been employed in Kott [3] and Garren and Chang [2]. In the latter paper, the Virginia telephone population is modeled as a Poisson sample from the entire population.

[5] Specifically, it is assumed that each $i \in \mathcal{U}$ has a probability $q_i > 0$ of being on the list $\mathcal{L}$ and that the $\Pr(i \in \mathcal{L} \text{ and } j \in \mathcal{L}) = q_i \cdot q_j$. Recognizing that $\mathcal{L}$ is not, in fact, a probability sample from $\mathcal{U}$, this assumption is a *model* for $\mathcal{L}$. The assumption that each q_i, for $i \in \mathcal{U}$, is positive implies that no subpopulation of $\mathcal{U}$ is systematically excluded from the list. It is permissible, for example, that purple polka dotted farmers have a low probability of making the list, but we assume that each purple polka dotted farmer could potentially be on the list. If a purple polka dotted farmer does make the list, he will receive a small value of q_i and hence a high undercoverage weight q_i^{-1} in equation (1.8) to represent all the purple polka dotted farmers that did not make the list.

A NML model, such as those developed here, produces estimates $\widehat{p}_i$ of p_i, and hence estimates $\widehat{q}_i = 1 - \widehat{p}_i$ of q_i. In particular if the $\widehat{\beta}_j$ are the fitted estimates from the NML model of the coefficients β_j in (1.3), then

$$\log\left(\frac{\widehat{p}_i}{1-\widehat{p}_i}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + ... + \widehat{\beta}_p X_{iP} \tag{1.10}$$

Hence redefining $\widehat{t}_{z0}$ as

$$\widehat{t}_z = \sum_{i\in\mathcal{L}} \frac{z_i}{\widehat{q}_i}, \tag{1.11}$$

we can use the NML model to correct the list population for undercoverage. NASS is currently using calibration for this purpose, where this model is implicit (see Kott [3]).

Assuming that the Poisson model for $\mathcal{L}$ is reasonable, and ignoring the fact that the $\widehat{q}_i$ are fitted probabilities, an estimate of the variance of $\widehat{t}_z$ is given by

$$\widehat{\mathrm{Var}}(\widehat{t}_z) = \sum_{i\in\mathcal{L}} (\widehat{q}_i^{-2} - \widehat{q}_i^{-1}) z_i^2. \tag{1.12}$$

Equation (1.12) gives the variability of $\widehat{t}_z$ under possible alternative realizations of $\mathcal{L}$ consistent with the Poisson model.

The perceptive reader will undoubtedly notice that since the NML model is developed using the area-frame survey, correction of the census for undercoverage, would require the X_{ij} be known for all $j \in \mathcal{L}$. In other words, variables used from the area frame survey for the purpose of developing the NML model must also be measured in the census.

One disadvantage of this procedure is that equation (1.10) develops undercoverage-corrected weights $\widehat{q}_i$ that can be unacceptably large. Section 1.4 discusses approaches to remedy this problem.

Let $\mathcal{L}_c$ and $\mathcal{U}_c$ be the list and target populations of some particular subpopulation c (for example, a specific county). Then equation (1.11) implies that

$$\widehat{N}_c = \sum_{i\in\mathcal{L}_c} \frac{1}{\widehat{q}_i}$$

estimates the total number of farms in the subpopulation. Thus if $N_{\mathcal{L}_c}$ is the number of list farms in the subpopulation c,

$$\widehat{N}_c - N_{\mathcal{L}_c} = \sum_{i\in\mathcal{L}_c} \left(\frac{1}{\widehat{q}_i} - 1\right) \tag{1.13}$$

estimates the number of NML farms in c. This estimate is probably smoother, less subject to sampling variation, than the design-based estimate of the number of NML farms in c computed from the AFS. The new estimate has great potential for allocating *future* area-frame surveys with the goal of "finding" more NML farms.

Alternatively if, for example, one is interested in improving the AFS estimate of the total number of NML horses in Wyoming, one can let z_i be the number of horses in the i-th list farm. Substituting into equation (1.11)

$$\sum_{i\in\mathcal{L}_c} z_i \left(\frac{1}{\widehat{q}_i} - 1\right) \tag{1.14}$$

estimates the amount of NML horses in the subpopulation c.

Equations (1.13) and (1.14) use an NML model to estimate NML counts and totals from census data. Since AFS stratum is a strong predictor of NML status but not available for census records, developing an alternative estimate for NML counts and totals using AFS rather than census data may prove fruitful. Again, letting c be a subpopulation (which can be defined in terms of AFS stratum), let $\mathcal{S}_c$ be the subsample of the AFS sample which lies in c. For $k \in \mathcal{S}_c$, let w_k be the sampling weight (expansion factor) and f_k be the adjusted tract-to-farm-acreage ratio. Then the estimators (1.13) and (1.11) can be replaced by

$$\sum_{k \in \mathcal{S}_c} w_k f_k \widehat{p}_k$$
$$\sum_{k \in \mathcal{S}_c} w_k f_k z_k \widehat{p}_k$$

respectively[6]. Again we expect these estimates to have less variance than simply looking at the NML farms in the AFS of c.

These formulae can be generalized to a list-based survey. Let $\tilde{\mathcal{L}}$ be a list for a NASS survey (which is genenally not identical to the full census mailing list), and suppose a sample $\tilde{\mathcal{S}}$ is taken from $\tilde{\mathcal{L}}$ with sampling probabilities $\tilde{\pi}_i$ and expansion factors $\tilde{w}_i = \tilde{\pi}_i^{-1}$. Suppose the members of $\tilde{\mathcal{L}}$ can be identified for all elements of the AFS. Then, using the techniques of this report, a model for $\tilde{p}_i$, the probability that the i-th farm is not in $\tilde{\mathcal{L}}$, can be developed.

Then equation (1.8) is replaced by

$$\widehat{t}_{z0} = \sum_{i \in \tilde{\mathcal{S}}} \tilde{w}_i \frac{z_i}{\tilde{q}_i}.$$

Estimating $\tilde{q}_i = 1 - \tilde{p}_i$ using the model, leads to the estimator

$$\widehat{t}_z = \sum_{i \in \tilde{\mathcal{S}}} \tilde{w}_i \frac{z_i}{\widehat{\tilde{q}}_i}$$

which replaces equation (1.11) when sampling is done from $\tilde{\mathcal{L}}$.

1.3 Experiments on coverage correction

As outlined in Section 1.2, one use of an NML model of the type developed here revolves around the correction of a list-based survey for undercoverage of the list. It proposes to make these corrections either on a large population (e.g. a state) for the direct purpose of estimation or on a smaller unit (e.g. a county) for the purpose of designing future area-frame surveys and supplements.

An experiment to assess the accuracy of such correction was performed as follows. The tracts in the California AFS were taken as the population $\mathcal{U}_A$. Tracts associated with farms on the mailing list were taken as the list population $\mathcal{L}_A$. The California model in equation (1.4) was used to generate $\widehat{q}_k = 1 - \widehat{p}_k$ for each $k \in \mathcal{U}_A$.

[6] There is a slight abuse of notation here. We will use i to index farms and k to index tracts. The AFS sampling unit is a tract; however the characteristics z of interest are farm level characteristics. Thus we will use z_i for the characteristic of the i-th farm and z_k for the characteristic of the farm which contains the k-th tract. We trust the context will make the meaning clear.

Table 1: "True" Population Values and List-based Coverage-corrected Estimates

x_i	t_z	$\widehat{t_z}$	stan err	t-ratio	
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	grow Christmas trees
CCENCATL *	14378.	14461.	1284.	0.065	raise cattle
ftypFRUT	29192.	29267.	974.	0.077	primary sales are fruit and nuts
ptfarm *	2164.0	2493.5	895.6	0.368	number of "point" farms
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	principal occupation of operator 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 stratum 11

Terms marked with an asterisk (*) do *not* appear in the model in equation (1.4)

For each $k \in \mathcal{U}_A$, let w_k be the sampling weight and f_k the adjusted tract-acreage-to-farm-acreage ratio. For various x_k as measured by the AFS (for example, CLANDTOT, the total acreage on the farm), let $z_k = w_k f_k x_k$. Thus

$$t_z = \sum_{k \in \mathcal{U}_A} z_k = \sum_{k \in \mathcal{U}_A} w_k f_k x_k$$

represents the AFS estimate of the CA total for the characteristic x_k.

For the purposes of this experiment, we treated t_z as the true population (that is $\mathcal{U}_A$) total of the z_k. We then pretended that the z_k for $k \in \mathcal{U}_A - \mathcal{L}_A$ and t_z were unknown and estimated t_z using the estimator $\widehat{t_z}$ of equation (1.11).

Representative selected results, out of the 69 x-variables considered, are listed in Table 1. To facilitate comparison between $\widehat{t_z}$ and t_z, the standard errors of $\widehat{t_z}$ under the Poisson model (see footnote 5 and equation (1.12)), as well as the resulting t-ratio's, are given. Of the 69 variables considered, the worst result (as measured by the t-ratio) was for the variable *strat11*, which is among the 14 shown on Table 1.

The conclusion of this experiment is that Poisson model for $\mathcal{L}_A$ and the fitted probabilities in equation (1.4) seem to fit. Moreover, equation (1.11) is a feasible correction for undercoverage.

1.4 Explorations of alternative link functions

Equation (1.3) implies the *link function*

$$p_i = (1 + e^{-\eta_i})^{-1}, \tag{1.15}$$

where

$$\eta_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{iP}. \tag{1.16}$$

In standard generalized linear modeling terminology, η is called the *linear link*.

In this study, we considered three other popular link functions[7]

$$\begin{aligned} p_i &= \Phi\left(\frac{\sqrt{2\pi}}{4}\eta_i\right) \\ p_i &= \exp\left(-\log(2)e^{-(2\log(2))^{-1}\eta_i}\right) \\ p_i &= 1 - \exp\left(-\log(2)e^{(2\log(2))^{-1}\eta_i}\right). \end{aligned} \tag{1.17}$$

These link functions are called the *probit*, the *loglog*, and the *complementary loglog* links respectively. As Figure 1 makes clear, these link functions are monotonically increasing, 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 loglog link dies much quicker for large negative values of η than it does for large positive values and the complementary loglog link dies quicker for large positive values of η than it does for large negative values.

The coverage experiments described in Section 1.3 were repeated for the link functions of equation (1.17). 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 in equation (1.4). 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. Since the interpretation of the coefficients β_i in terms of the odds ratio, see equation (1.5), only applies to the logistic link, and since the logistic link leads to a slightly simpler fitting algorithm, the logistic link is the only link explored for the remainder of this paper.

We did not investigate the *inverse-linear* link function used implicitly by NASS for calibration. It has the form:

$$p_i = (1 + \eta_i)^{-1} \tag{1.18}$$

When using calibration to correct for undercoverage, current NASS integerized the calibration weights limited them to a minimum of 1 and a maximum of 6.

Table 3 shows the effects, separately, of integerizing and truncating the implied undercoverage weights $(1 - \widehat{p}_i)^{-1}$ based on the logistic link. We found that the truncation at 6 (truncation from below is unnecessary with a logistic link) does not have a sizeable effect on the quality of the undercoverage correction, but that the integerization substantially degrades it.

[7]For comparison purposes, the link functions have been normalized so that $p = 0.5$ and $\frac{\mathrm{d}p}{\mathrm{d}\eta} = 0.25$ when $\eta = 0$.

Link Functions

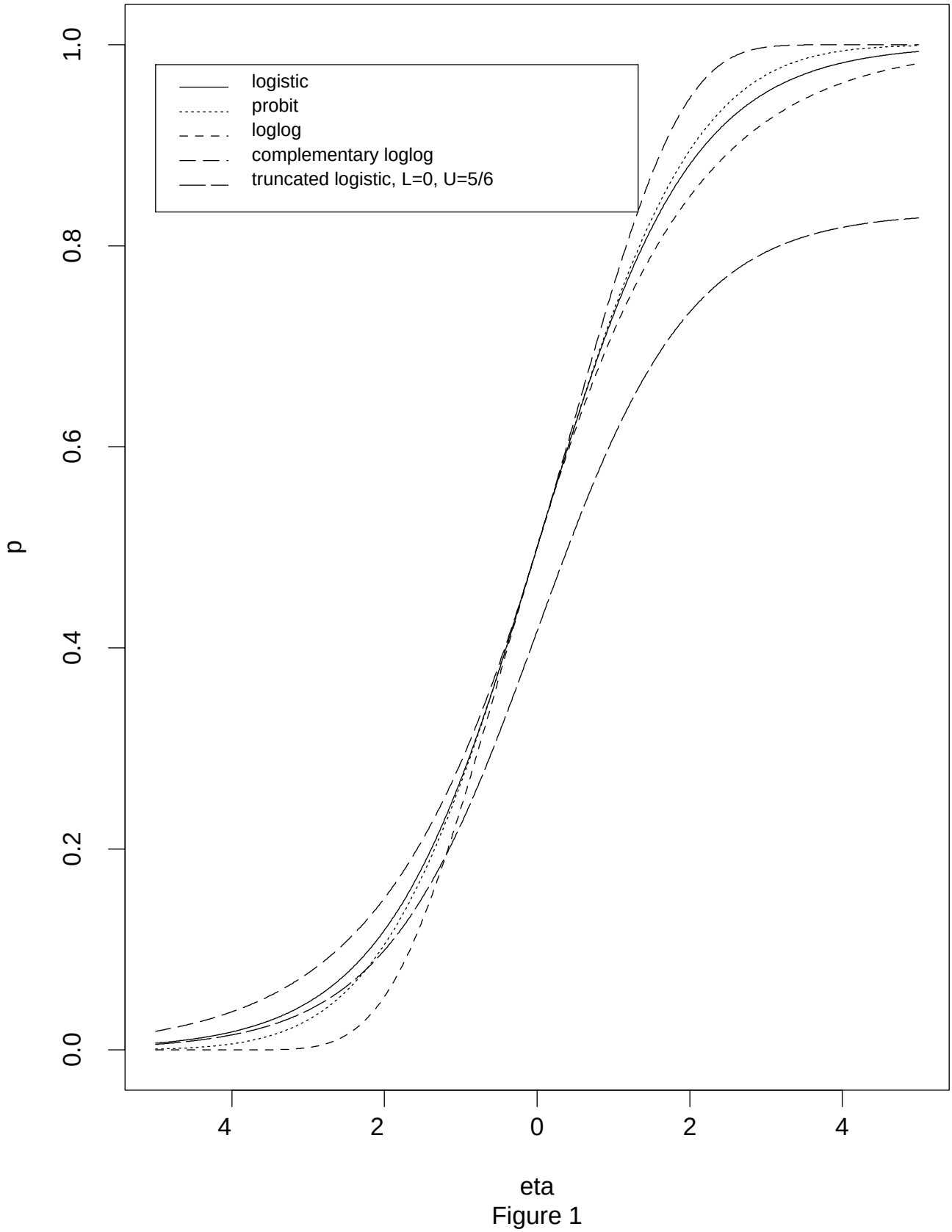

Figure 1

Table 2: Alternative Link Functions

x_i	t_z	logistic	probit	loglog	complementary loglog
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
CCENCATL	14378.	14461.	14368.	14388.	14393.
ftypFRUT	29192.	29267.	29315.	29450.	29172.
ptfarm	2164.0	2493.5	2240.1	2369.4	2482.8
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.

Folsom and Singh [1] propose the link function[8]

$$p_i = L + (U - L)(1 + \exp(-\eta_i))^{-1}. \tag{1.19}$$

Since $(1-p)^{-1} < 6$ is equivalent to $p < 5/6$, one can use $L = 0$ and $U = 5/6$ to limit the undercoverage weights to 6. Table 3 shows the results from using this link. It appears to provide no advantage over simply truncating the weights from the logistic link.

2 The California Data

The predictor variables were divided into 7 groups as follows:

sales (10 variables). This consists of the variables *sales1K*, *sales2.5K*, *sales5K*, *sales10K*, *sales25K*, *sales50K*, *sales100K*, *sales250K*, *sales500K*, *sales1000K*, where, for example, *sales5K* = 1 when annual sales is at least \$5000 and 0 otherwise. This group is a recode of MFARMDEF.

ptfarm (1 variable): *ptfarm* = 1 - *sales1K*. This variable is 1 when the farm is classified as a farm using the point system and does not have \$1,000 in sales. It is 0 otherwise.

ftype (48 variables). The AFS has 16 categories of farm products: GRAN (grains), TOBA (tobacco), COTT (cotton), VEGM (vegetables), FRUT (fruit and nuts), NURS (nursery), CHRS (Christmas trees), OTHC (other crops and hay), HOGS, MILK, CATL (cattle), SHEP (sheep), EQUI (horses), POUL (poultry), AQUA (aquaculture), and OTHA (other animal products). *For each of these 16 groups*, for example GRAN, three variables were created:

[8] When an intercept is included in the model, equation (1.19) is algebraicly equivalent to weight function given in Folsom and Singh [1].

Table 3: Truncating and Integerizing Weights

x_i	t_z	logistic link	logistic link truncated weights	logistic link integerized weights	truncated logistic link
1	68896.	68727.	68338.	65084.	68005.
CCENGRAN	5701.3	5882.3	5882.3	5577.1	5931.9
CCENCHRS	299.20	190.17	190.17	241.53	197.63
CCENCATL	14378.	14461.	14461.	14087.	14533.
ftypFRUT	29192.	29267.	29267.	26715.	29283.
ptfarm	2164.0	2493.5	2493.5	2429.1	2484.7
sales10K	39564.	38371.	38371.	35014.	38441.
sales100K	19912.	19576.	19576.	18300.	19637.
sales1000K	5860.6	5884.8	5884.8	5731.8	5882.8
hisp	8088.8	7844.0	7730.0	7574.8	7390.2
ocup	37002.	38343.	38229.	36441.	38364.
LEQUIOWN	87805.	78990.	78876.	77627.	78468.
CLANDTOT	22962.	23375.	23368.	21420.	23489.
strat11	17268.	19540.	19265.	18558.	19109.

CCENGRAN = 1 when the farm had this type of income, 0 otherwise;
ftypGRAN =1 when the primary source of income was from this category, 0 otherwise (recode of MFARMTYP); and
GRAN = CCENGRAN - ftypGRAN.

There were no tobacco farms in the CA AFS.

oper (eight variables). These variables describe characteristics of the prinicpal operator.

- *age*: the age of principal operator to the nearest multiple of 10 years. The possible values of age are 20, 30, 40, 50, 60, 70 (recode of MDEMOAGE).
- *gender*: 1 when male, 0 when female (recode of MDEMOSEX)
- *hisp*: 1 when Spanish, Hispanic, or Latino origin, 0 otherwise (recode of MDEMHISP).
- *black*, *indian*, *hawaii*, *asian*: 1 when reported given race, 0 otherwise (recode of MDEMRACE).
- *ocup*: 1 when the principal occupation of the principal operator is farming or ranching, and 0 otherwise (recode of MDEMOCUP).

There were no reported Native Hawaiians in the California AFS.

horse (seven variables). These variables refer to equine ownership.

- LEQUIOWN: the number of horses and ponies owned.
- LEQUOTOW: the number of mules, donkeys, or burros owned.

- *leqoper1*, *leqoper2*, *leqoper3*, *leqoper4*, *leqoper5* (recodes of LEQUOPER). These indicator variables are 1 when, respectively, the operation is a farm or ranch (1), a boarding, training or riding facility (2), a breeding service place (3), a place to keep equine for personal use (4), or some other type of operation with equine (5).

Due to low counts, in analyzing the California data set, *leqoper2* and *leqoper3* were combined, and *leqoper5* was deleted.

land (three variables). These variables refer to land acreage.

- CLANDTOT: total acres operated.
- CROPLAND: total acres of cropland.
- CLANDCRP: total acreage in the Conservation or Wetland Reserve Programs.

stratum (nine variables). The AFS in California has 10 strata, numbered 11, 17, 19, 21, 27, 31, 32, 41, 45, and 50. The variables - *strat11*, *strat17*, *strat19*, *strat21*, *strat27*, *strat31*, *strat32*, *strat41*, and *strat45* - was set to 1 when the farm was in the indicated stratum and to 0 otherwise. Although a sample of segments was taken from Stratum 50, none of those segments included a tract from a farm. In the Illinois/Indiana/Iowa AFS, the strata are defined differently and indicator variables were created accordingly. The strata variables for the 48-state sample were constructed differently as described in Section 4.

In the data sets we analyzed, there were no missing values among these variables for the tracts identified as parts of farms.

Tracts were weighted using a product of a sampling weight (MCOMBADJ) and a tract-acreage-to-farm-acreage ratio (coded "weight" in Herb Eldridge's extraction routine, Section 8.1), modified slightly to account for the space occupied by a potential house.

A modified stepwise regression procedure was used for variable selection. This procedure is described in greater detail in Section 5.2. Starting with a base model with an intercept only, each group of variables was separately tested to see whether the addition of the group represented a significant improvement over the base model. The most significant group was selected and a stepwise regression procedure used to decide which variables within the group were needed. The resulting model became the new base model, and the procedure was iterated. All significance tests were performed using design-based linearization variance estimates for logistic regressions in sample surveys.

Based upon significance, the groups that were entered into the model for Califonia were sales, ftype, and oper, in that order.

The stratum variables were handled differently. Because AFS stratum is not an available variable for Census records, a model using AFS stratum can not be employed to correct the Census for undercoverage. In order to determine how much NML can be predicted using variables not related to AFS stratum, the stratum group was not entered into the model until no other group not already entered into the model was significant.

Table 4 presents two models, one with variables from the stratum group and one, displayed in equation (1.4), without.

Table 4: California NML Model With and Without Strata

effect	without stratum model coefficient (stan. err.)	with stratum model coefficient (stan. err.)
1	2.4417 (0.5534)	2.6810 (0.5876)
sales5K	-1.0351 (0.2130)	-0.9776 (0.2166)
sales50K	-0.8126 (0.2590)	-0.8327 (0.2638)
sales1000K	-1.7876 (0.8203)	-1.7313 (0.8197)
CHRS	-1.2506 (0.3115)	-1.7968 (0.3344)
CCENFRUT	-0.9089 (0.2100)	-0.9604 (0.2132)
CCENCOTT	-2.6183 (1.0192)	-2.1375 (0.9821)
CCENSHEP	0.9662 (0.4163)	0.8404 (0.4302)
CCENAQUA	2.1035 (0.8686)	2.1269 (0.8127)
age	-0.03621 (0.00960)	-0.03850 (0.0100)
hisp	1.1397 (0.3032)	1.0183 (0.3011)
asian	1.0276 (0.3305)	1.0177 (0.3159)
ocup	-0.5709 (0.2700)	-0.5763 (0.2789)
strat32		5.8508 (1.0917)
strat45		1.0244 (0.2880)
strat11		-0.6283 (0.2559)

2.1 The nursery in Stratum 32 and its implications

The California AFS has one farm in Stratum 32 ("dense urban: over 100 homes per square mile"): a nursery with sales between $500,000 and $1,000,000 operated by a 40-year-old Hispanic. This nursery happens not to be on the list. Using the coefficients in Table 4 for the model with stratum effects, the fitted probability that this farm is not on the list is

$$(1 + exp[-(2.6810 - 0.9776 - 0.8327 - 0.03850 * 40 + 1.0183 + 5.8508)])^{-1} = 0.998.$$

Notice that in the absence of the term for Stratum 32, the fitted probability would have been

$$(1 + exp[-(2.6810 - 0.9776 - 0.8327 - 0.03850 * 40 + 1.0183)])^{-1} = 0.586.$$

The highly significant coefficient of 5.8508 (standard error 1.0917) improves the fit for the solitary nursery in Stratum 32 without degrading the fit for any other farm in the sample (since there are no other farms in Stratum 32). The design-based linearization standard error estimates are based upon an asymptotic large sample approximation. Although the total sample size for the CA sample is 1,488, this particular coefficient is due to only a single farm. One needs to interpret the coefficient with extreme caution.

Consider the nine stratum zero-one variables *strat11*, *strat17*, *strat19*, *strat21*, *strat27*, *strat31*, *strat32*, *strat41*, *strat45*. Since each farm is in exactly one stratum

$$\begin{aligned} 1 &= \text{strat11} + \text{strat17} + \text{strat19} + \text{strat21} + \text{strat27} \\ &+ \text{strat31} + \text{strat32} + \text{strat41} + \text{strat45}. \end{aligned} \tag{2.20}$$

Suppose a model were developed with a complete set of stratum variables. Due to equation (2.20), only eight of the stratum variables would be needed to completely specify the model.

If the variable *strat32* were not used, then

$$\text{strat11} + \text{strat17} + \text{strat19} + \text{strat21} + \text{strat27} + \text{strat31} + \text{strat41} + \text{strat45}$$

would be a highly significant linear combination of variables whose significance, in fact, would only be due to a single data point.

In other words, it is possible that the analysis has a highly significant hidden linear combination of variables which, in fact, is due to only a small number of data points. Developing methods for detecting the presence this phenomenon is a pressing methodological problem in need of further research.

3 The Illinois/Indiana/Iowa Data

Herb Eldridge has provided, in the same format as the California data set, the AFS for Illinois, Indiana, and Iowa. The same variables were used except that there is now an additional group of indicator variables for state. In addition, the AFS stratum group has ten variables. There is a total of 5,841 farms in this data set, almost four times as many as in the California data set. Thus, one would expect that more variables would be significant than in the California model.

As in the California data set, the most important group was sales, followed by farm type and operator characteristics. These are the same groups uncovered for the California data set, but the order of farm type and operator characterics are reversed. The fourth group to enter was the land (land-acreage) group. Stratum was also significant, but by fiat for the reasons outlined earlier, it was used last. Interestingly, the state group of variables was not significant.

Table 5 has the fitted models, with and without the stratum variables.

Examining Tables 4 and 5, besides the importance of sales, one sees that age and occupation of the principal operator are significant predictors. The sign of *asian* changes dramatically between the two models. This is likely because there is only one Asian in the Ilinois/Indiana/Iowa AFS, another apparent version of the solitary-nursery problem noted in Section 2.1.

The authors have no explanation for the apparent change is sign of the coefficient of *hisp*. Perhaps the population of Hispanic farmers is different in California from that in Illinois/Indiana/Iowa in ways not captured by the other variables in the model(s). Christmas trees remain important in both models. So do aquaculture and fruits and nuts, both of which have coefficients that change sign. One should be aware that there are only three aquaculture farms in the Illinois/Indiana/Iowa data set and that the coefficient of fruit and nuts, in fact, ceases to be significant when *strat33* ("resort: over 20 homes per square mile") enters the model.

Table 5: Illinois/Indiana/Iowa NML Model With and Without Strata

effect	without stratum model coefficient (stan. err.)	with stratum model coefficient (stan. err.)
1	0.5078 (0.3658)	0.5389 (0.3660)
sales2.5K	-1.0820 (0.2231)	-1.1358 (0.2154)
sales10K	-0.5833 (0.2266)	-0.5625 (0.2257)
sales100K	-0.5906 (0.2832)	-0.6233 (0.2814)
sales1000K	-1.6463 (0.5951)	-1.6708 (0.5929)
ftypCHRS	-6.5169 (0.5220)	-6.5204 (0.5225)
ftypVEGM	-5.3687 (0.4252)	-5.3186 (0.4180)
CCENEQUI	0.6347 (0.1726)	0.5673 (0.1731)
OTHA	-5.8550 (0.4445)	-5.8480 (0.4420)
AQUA	-2.7365 (1.0579)	-2.7036 (1.0601)
ftypNURS	1.4875 (0.5817)	1.5029 (0.5799)
ftypFRUT	2.2507 (0.9370)	1.0425 (0.7253)
asian	-4.3986 (1.0325)	-4.3407 (1.0319)
ocup	-0.5567 (0.1747)	-0.5150 (0.1735)
age	-0.01442 (0.00667)	-0.01485 (0.00669)
hisp	-1.2444 (0.6475)	-1.2319 (0.6486)
CLANDCRP	-0.02996 (0.00837)	-0.02943 (0.00837)
CROPLAND	-0.0007010 (0.0003431)	-0.0007161 (0.0003476)
strat33		4.2082 (0.5503)

4 The 48-State Model

There are 46,000 farms (exactly!) in the 48-states AFS. The variables as defined in Section 2 were changed somewhat as follows.

strat. The strata definitions are slightly different from state to state. On Bill Wigton's advice, variables strat10s, strat20s etc. were created where, for example, strat10s = 1 when the farm is in a stratum numbered between 10 and 19. The group now has five variables, including strat50s60s which covers strata numbered between 50 and 69.

oper. The gender variable was changed to a variable *female* that is 1 when the principal operator is female and 0 otherwise. The age variable was changed to $agec = (age - 45)/10$.

land. Two variables were added: CLAND519 (percentage change in total market value of all land and buildings since June 1, 2001) was added. CLAND518 was recoded to *clandval* which takes the values +1, -1, and 0 if the total market value of all land and buidlings increased, decreased, or was without change.

The groups in order of entry were sales, operator characteristics, land acreage, farm type, state, type of equine operation, and finally, by fiat, stratum. At each stage of the selection process, the most significant (except for stratum) of the remaining groups was next selected for consideration.

Examining Table 6, the fitted coefficients exhibit great stability: their values are in a believable range (see the discussion on numerical stability in Section 5.2), and they do not change much even when additional terms are entered into the model. There appears to be much benefit from increasing the sample size by pooling data across states, even when the individual states appear to have a great deal of data.

4.1 The main effects model

Table 6 gives the 48-state models for models with main effects only. Because of the much larger size of the data, more terms are significant. Two models are given: one with stratum variables and one without. In addition, the coefficients for the main effects in a model with interaction terms are given in Table 6. The coefficients of the interaction terms are given in Table 7.

4.2 The model with interactions

Due to time limitations, the only interactions that were fit were between state and the other variables.

Suppose X_1 and X_2 are two variables. The "interaction" of X_1 and X_2 is the product variable $X_1 \cdot X_2$. If $X_{+11}, \cdots, X_{+1u}$[9] code the u levels of a categorical variable A and $X_{+21}, \cdots, X_{+2v}$ code the v levels of a categorical variable B, the uv variables $X_{+1r} \cdot X_{+2s}$ code the interaction A $*$ B. For example, in this study, state is coded in 48 variables and ftype in 16 variables so state $*$ ftype is coded in 768 $(= 48 \cdot 16)$ variables.

[9]It is convenient here to double subscript all variables. Thus specific observations of these variables are triple subscripted with the first subscript indicating observation number and the latter two subscripts indicating variable.

Table 6: The 48-State Model With and Without Strata
(including the main effects for the model with interactions)

effect	without stratum model coefficient (stan. err.)	with stratum model coefficient (stan. err.)	interactions model (main effects)	interactions model (main effects) no strata variables
1	0.3087 (0.0655)	0.4484 (0.0739)	0.3688 (0.0565)	0.3016 (0.0534)
sales2.5K	-0.4761 (0.2231)	-0.4668 (0.0575)	-0.4623 (0.0576)	-0.4580 (0.0575)
sales5K	-0.2069 (0.0722)	-0.2033 (0.0720)	-0.1679 (0.0724)	-0.1842 (0.0724)
sales10K	-0.3597 (0.2266)	-0.3562 (0.0866)	-0.3508 (0.0880)	-0.3416 (0.0882)
sales25K	-0.2798 (0.0968)	-0.2757 (0.0971)	-0.2484 (0.0977)	-0.2795 (0.0980)
sales50K	-0.2600 (0.1121)	-0.2566 (0.1124)	-0.2904 (0.0992)	-0.3072 (0.0992)
sales100K	-0.2933 (0.1264)	-0.2860 (0.1265)		
sales250K	-0.7504 (0.1613)	-0.7174 (0.1606)	-0.7349 (0.1494)	-0.7592 (0.1494)
agec	-0.02180 (0.00170)	-0.02201 (0.00170)	-0.02164 (0.00171)	-0.02116 (0.00170)
ocup	-0.2806 (0.0486)	-0.2786 (0.0488)	-0.2713 (0.0488)	-0.2741 (0.0488)
female	0.2925 (0.0667)	0.2841 (0.0673)	0.2998 (0.0668)	0.2934 (0.0671)
black	1.0373 (0.1560)	1.0198 (0.1558)	1.0255 (0.1553)	1.0429 (0.1560)
hisp	0.3946 (0.0958)	0.4155 (0.0969)	0.4789 (0.0946)	0.4464 (0.0945)
asian	0.6447 (0.2872)	0.7135 (0.2867)	0.6020 (0.2803)	0.6022 (0.2854)
CLANDTOT	-.00006657 (.00001479)	-.00008954 (.00001571)	-.0008502 (.0000744)	-.0007905 (.0000722)
CLANDCRP	-0.009039 (0.001285)	-0.008714 (0.001273)	-0.01813 (0.00222)	-0.01898 (0.00223)
OTHC	-0.4248 (0.0477)	-0.4320 (0.0476)	-0.3734 (0.0471)	-0.3599 (0.0472)
ftypOTHC	-0.1755 (0.0731)	-0.2007 (0.0730)		
ftypEQUI	0.4231 (0.0797)	0.3896 (0.0798)	0.4302 (0.0728)	0.4351 (0.0731)
CCENGRAN	-0.3466 (0.0707)	-0.3189 (0.0716)	-0.3389 (0.0582)	-0.3895 (0.0576)
ftypGRAN	-0.2102 (0.0943)	-0.1969 (0.0946)		
CCENNURS	0.6223 (0.1627)	0.5930 (0.1629)	0.5519 (0.1581)	0.5802 (0.1589)
HOGS	0.4768 (0.1101)	0.4742 (0.1096)	0.4155 (0.1101)	0.4503 (0.1103)
CCENCATL	-0.3048 (0.0589)	-0.3493 (0.0596)	-0.2768 (0.0505)	-0.2571 (0.0504)
CCENCOTT	-0.5097 (0.1618)	-0.4271 (0.1623)	-0.3855 (0.1605)	-0.4332 (0.1613)
ftypTOBA	-0.9373 (0.2043)	-0.9764 (0.2069)	-0.9699 (0.2019)	-0.9321 (0.2012)
ftypCHRS	0.4278 (0.2148)			
IL	-0.4565 (0.1332)	-0.3760 (0.1329)	-0.4172 (0.1333)	-0.4741 (0.1323)
IN	-0.3937 (0.1404)	-0.3144 (0.1405)	-0.3998 (0.1392)	-0.4368 (0.1382)
IA	-0.4825 (0.1265)	-0.3860 (0.1260)	-0.3998 (0.1264)	-0.4680 (0.1261)
MN	-0.2607 (0.1218)		-0.3747 (0.1306)	-0.2457 (0.1233)
MS	0.3283 (0.1501)	0.3190 (0.1482)	0.7554 (0.2114)	0.7878 (0.2117)
MO	0.3080 (0.0876)	0.2976 (0.0886)	0.3204 (0.0881)	0.3206 (0.0874)
NE	-0.6608 (0.1762)	-0.5445 (0.1749)	-0.7081 (0.1854)	-0.8048 (0.1833)
NH			0.5257 (0.2305)	0.5792 (0.2301)
RI	1.9015 (0.6438)	1.9032 (0.6097)	2.4427 (0.5443)	2.4378 (0.5410)
SD	-0.5316 (0.2345)			
TX	0.2451 (0.0706)	0.2417 (0.0713)		
VT	0.8078 (0.2796)	0.7812 (0.2810)	1.0455 (0.3886)	1.0465 (0.3817)
leqoper4	0.4337 (0.0740)	0.4380 (0.0741)	0.4294 (0.0729)	0.4249 (0.0736)
leqoper5	0.8010 (0.2094)	0.8071 (0.2066)	0.7410 (0.2042)	0.7913 (0.2098)
strat10s		-0.2982 (0.0587)	-0.2298 (0.0486)	
strat20s		-0.1236 (0.0561)		

The model

$$\eta_i = \beta_0 + \beta_{11}X_{i11} + \cdots + \beta_{1u}X_{i1u} + \beta_{21}X_{i21} + \cdots + \beta_{2v}X_{i2v}$$

is referred to as a *main effects* model, and the model

$$\begin{aligned} \eta_i &= \beta_0 + \beta_{11}X_{i11} + \cdots + \beta_{1u}X_{i1u} + \beta_{21}X_{i21} + \cdots + \beta_{2v}X_{i2v} \\ &+ \gamma_{11}X_{i11} \cdot X_{i21} + \cdots + \gamma_{rs}X_{i1r} \cdot X_{i2s} + \cdots + \gamma_{uv}X_{1u} \cdot X_{2v} \end{aligned} \tag{4.21}$$

is traditionally called a model *with interactions*.

Equation (4.21) has too many parameters. As a result, it is conventional to put restrictions on parameter values to ensure a unique solution. Our model fitting techniques, however, were fairly conservative, especially with interaction coefficients γ_{rs}. Only the most significant ones were entered into the model. The rest were implicitly set to 0. As a consequence, (further) restrictions on parameter values were unnecessary.

Another common practice is to allow $\gamma_{rs} \neq 0$ only when both $\beta_{1r} \neq 0$ and $\beta_{2s} \neq 0$. We have are only fitting interactions between states and the other variables, however. The main effects represent an overall national model, and the few interactions that we fit represent significant state-level deviations from the national model. We estimated the overall coefficient for North Dakota, say, to be 0, signifying that there was no greater (or lesser) systematic tendency for North Dakota farms to be NML, while the estimated coefficient of ND.leqoper1 was negative, signifying that farms or ranches with horses in North Dakota were less likely to be NML than farms or ranches with horses in other states. states. In our modeling, we allowed $\gamma_{rs} \neq 0$ even when both β_{1r} and β_{2s} fitted to zero.

Suppose $X_{+11}, \cdots, X_{+1u}$ code the u levels of a categorical variable A (for example state) and X_2 represents a continuous variable (such as CLANDTOT - total acreage), the main effects model looks like

$$\eta_i = \beta_0 + \beta_{11}X_{i11} + \cdots + \beta_{1u}X_{i1u} + \beta_2 X_{i2}$$

and the interactive model looks like

$$\eta_i = \beta_0 + \beta_{11}X_{i11} + \cdots + \beta_{1u}X_{i1u} + \beta_2 X_{i2} + \gamma_1 X_{i11} \cdot X_{i2} + \cdots + \gamma_u X_{i1u} \cdot X_{i2}.$$

In this case, β_2 represents a slope and, when the categorical variable A is state, the γ_r represent the difference of a specific slope for the r-th state from the overall national slope.

As discussed in Section 5.2, the models became numerically unstable as soon as many interactions were fit. Thus, the approach we used for allowing non-zero interaction parameters in the model was extremely conservative.

Examining Table 7, one notices immediately that interactions between the states and the acreage variables (CLANDTOT, CROPLAND, CLANDCRP) were far more numerous than interactions between the states and other types of variables. One possible explanation is the large variation in farm sizes, as measured by acreage, among the various states. A second, somewhat more mysterious observation, is that although the variables relating to changes in value of land and buildings (CLAND519, *clandval*) were not predictive of NML status nationally, they were predictive in several New England states (and only in New England states).

Tables 6 and 7 also include a "no stratum variables" model with interactions. The model fitting procedure was not applied to select variables for this model. Rather the variables from the "with stratum variables" model with interaction were chosen and the model refit with all stratum variables removed.

Table 7: Interaction Terms for the 48-State Models

effect	with strata variables coefficient (stan. err.)	without strata variables coefficient (stan. err.)
MI*sales500K	-2.1177 (0.7777)	-2.1430 (0.7786)
WY*hisp	-2.6698 (0.4467)	-2.5872 (0.4420)
MS*CCENOTHC	-1.0766 (0.2857)	-1.0672 (0.2851)
SC*CCENVEGM	-3.1860 (0.9773)	-3.1314 (0.9802)
TX*CCENMILK	-5.8064 (0.6992)	-5.7543 (0.6962)
AZ*CLANDTOT	-0.004782 (0.0000744)	-0.004319 (0.0000722)
AZ*CROPLAND	0.005751 (0.0002220	0.005179 (0.000268)
AR*CLANDCRP	0.01852 (0.00311)	0.01893 (0.00392)
CA*CLANDCRP	0.01840 (0.00224)	0.01927 (0.00225)
CA*CLANDTOT	0.0008535 (0.0000750)	0.0007966 (0.0000726)
CO*CLANDCRP	0.01526 (0.00243)	0.01592 (0.00245)
CO*CLANDTOT	0.0008334 (0.0000834)	0.0007813 (0.0000800)
GA*CLANDCRP	-0.2035 (0.0035)	-0.2009 (0.0035)
ID*CLANDCRP	-0.01798 (0.00316)	-0.01863 (0.00318)
MT*CLANDCRP	0.01362 (0.00298)	0.01410 (0.00302)
NE*CLANDTOT	0.0007838 (0.0001163)	0.0007573 (0.0001024)
NH*clandval	-1.7723 (0.2693)	-1.7996 (0.2719)
NH*CLAND519	-0.1072 (0.0158)	-0.1050 (0.0162)
NM*CLANDTOT	0.0008287 (0.0000816)	0.0006988 (0.0000901)
OK*CLANDCRP	0.01928 (0.00247)	0.02002 (0.00253)
RI*clandval	-2.3268 (0.5428)	-2.4841 (0.5572)
TX*CLANDCRP	0.01457 (0.00246)	0.01516 (0.00248)
TX*CLANDTOT	0.0008511 (0.0000743)	0.0007938 (0.0000721)
UT*CLANDCRP	0.01935 (0.00229)	0.01992 (0.00293)
UT*CLANDTOT	0.0003800 (0.0000987)	0.0003399 (0.0000982)
VT*clandval	-2.7349 (0.6306)	-2.7019 (0.6073)
VT*CLAND519	-0.3138 (0.0580)	-0.3201 (0.0605)
WA*CLANDCRP	0.01872 (0.00245)	0.01938 (0.00249)
WY*CLANDTOT	0.0007227 (0.0001442)	0.0006776 (0.0001369)
GA*leqoper5	4.3925 (0.5976)	4.1552 (0.6114)
ND*leqoper1	-1.8738 (0.6708)	-1.9777 (0.6736)
AZ*strat30s	12.4190 (0.6427)	
MN*strat40s	1.8066 (0.3887)	
NM*strat40s	-0.6421 (0.1579)	
NC*strat30s	-14.0901 (0.9916)	
WI*strat30s	-11.9179 (0.8102)	

4.3 Interpreting the results

What exactly do Tables 6 and 7 tell us about the nature of 2002 NML farms in the 48 contiguous states? As expected, farms with larger sales were less likely to be NML. Similarly, farms with any cotton, cattle, grain crops, or other crops and hay were *less* likely to be on the NML, all other things being equal, as were farms with grains crops, tobacco, or other crops and hay as their primary source of sales. By contrast, farms with equine or Christmas trees as their primary source of sales or *any* nursery items were *more* likely to be NML, all other things being equal. One small surprise: farms with hogs as their primary source of income were *not* significantly more likely to be on the Census Mail list, while farms with hogs that were not primarily hog farms were *more* likely to be NML than farms that were otherwise just like them.

Even taking sales size, farm type, and the age of the principal operator into consideration, farms operated by blacks, Asians, Hispanics, and women were more likely to be NML. The estimated coefficient on blacks is particularly striking.

Surprisingly, point farms were not more likely to be NML than other farms with less than $2,500 in annual sales. This may be because many point farms were so designated because they had horses, and an operation (as defined by the Census process) having horses for either personal use or another unspecified use (not boarding, training, riding, or breeding) was more likely to be NML than a farm otherwise just like it.

Farms in Mississippi, Missouri, Texas, and some of the New England states were more likely to be NML than in other states, all other things being equal. Farms in the AFS Strata 10 to 29 were less likely to be NML, which means that farms in AFS Strata 30 to 69 were more like to be NML. Recall that this is after accounting for size and farm type. In Arizona, farms in the 30's were particularly likely to be NML, while farms in the 40's in neighboring New Mexico were *less* likely to be NML all other things being equal. Other notable positive interaction coefficients were for other-use equine-owning operations in Georgia and for Strata-40-to-49 farms in Minnesota. Farms in the 30's in both North Carolina and Wisconsin, Hispanics in Wyoming, and farms with increasing land values in some New England states, by contrast, were more likely *not* NML than farms otherwise just like them.

5 Statistical Methodology

5.1 Calculation of standard errors and significance tests

In the AFS, there are three levels of stratification: state, stratum within state, and substratum. For the purpose of analysis, it suffices to use a single index h for the combined stratification, and we will refer to stratum h when we really mean a specific substratum of a particular stratum in a certain state.

We will use the index r to denote primary sampling units (PSU's), which are the area-sample segments for our purposes. We assume that the PSU's are chosen *with replacement* (more on this assumption later). We will refer to the r-th segment in the h-th stratum as 'the hr-th segment'.

We will use the index k to denote tracts within segments. All the characteristics we are measuring are farm level characteristics. Thus we will use $\mathbf{z}_{hrk}$ to denote a vector of characteristics of the farm which contains the k-th tract in the hr-th segment[10]. In what

[10]In earlier sections, where issues of the sampling design did not arise, we used a single index k to index

follows all vectors are assumed to be column vectors.

We will continue to use the index i for farms. Thus, for example, X_{ij} represents the value of the j-th variable on the i-th farm, and X_{hrkj} represents the value of the same variable on the farm which contains the hrk-th tract.

Let w_{hr} denote the sampling weight (expansion factor) of the hr-th segment and let f_{hrk} denote the tract-acreage-to-farm-acreage ratio associated with the hrk-th tract. Let $\mathcal{U}_h$ and $\mathcal{S}_h$ be the population and sample of segments in stratum h. Then the population total of $\mathbf{z}$ is given by

$$t_{\mathbf{z}} = \sum_{i \in \mathcal{F}} \mathbf{z}_i = \sum_h \sum_{r \in \mathcal{U}_h} \sum_k f_{hrk} \mathbf{z}_{hrk},$$

where $\mathcal{F}$ is the population of farms and for $i \in \mathcal{F}$, $\mathbf{z}_i$ is the vector of characteristics on farm i. Thus $t_{\mathbf{z}}$ can be unbiasedly estimated by

$$\widehat{t}_{\mathbf{z}} = \sum_h \sum_{r \in \mathcal{S}_h} \sum_k w_{hr} f_{hrk} \mathbf{z}_{hrk}.$$

Since each segment in $\mathcal{S}_h$ is completely subsampled, we suppress the dependence of k on hr in our formulae.

Letting

$$\begin{aligned} \mathbf{z}_{hr} &= n_h \sum_k w_{hr} f_{hrk} \mathbf{z}_{hrk} \\ \overline{\mathbf{z}}_h &= n_h^{-1} \sum_{r \in \mathcal{S}_h} \mathbf{z}_{hr}, \end{aligned}$$

where n_h is the number of segments in $\mathcal{S}_h$, we have

$$\widehat{t}_{\mathbf{z}} = \sum_h \overline{\mathbf{z}}_h. \tag{5.22}$$

Since the sampling of the PSU's is (treated as if it were) with replacement, for each h, the $\mathbf{z}_{hr}$ are independent and identically distributed random variables. An unbiased estimator for the variance of $\widehat{t}_{\mathbf{z}}$ is consequently

$$\widehat{Var}(\widehat{t}_{\mathbf{z}}) = \sum_h n_h^{-1} \frac{\sum_{r \in \mathcal{S}_h} (\mathbf{z}_{hr} - \overline{\mathbf{z}}_h)(\mathbf{z}_{hr} - \overline{\mathbf{z}}_h)^T}{n_h - 1}. \tag{5.23}$$

.

A design-based approach for generalized linear modeling in the sample survey setting is described, somewhat abstractly, in [6] Section 13.4. A good reference for generalized linear modeling in the standard statistical context is [5]. We outline this approach below with slight modifications that apply to the NASS AFS design.

Consider the superpopulation model

$$y_i \text{ is distributed binomial}(1, p_i), \quad i \in \mathcal{F} \tag{5.24}$$

$$y_1, \cdots, y_N \text{ are independent}$$

$$p_i = g(\eta_i) \tag{5.25}$$

$$\eta_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{iP}. \tag{5.26}$$

tracts. In this sections, tracts are specified by a triple index hrk.

In the context of this report $y_i = 1$ if the i-th farm has NML status and 0 otherwise. Let $\mathbf{X}_{i+}$ be the $(P+1)$-length vector

$$\mathbf{X}_{i+} = [\ 1 \quad X_{i1} \quad \cdots \quad X_{iP}]^T.$$

For the superpopulation model in equation (5.24), the maximum likelihood estimate $\mathbf{B}$ of $\beta = [\beta_0 \quad \cdots \quad \beta_P]^T$ satisfies

$$\mathbf{L}(\mathbf{B}) = \mathbf{0},$$

where $\mathbf{L} : R^{P+1} \to R^{P+1}$ is the map

$$\mathbf{L}(\beta) = \sum_{i \in \mathcal{F}} \frac{y_i - p_i}{p_i(1-p_i)} g^{'}(\eta_i)\mathbf{X}_{i+}, \tag{5.27}$$

and $g^{'}(\eta_i) = dg(\eta_i)/d(\eta_i)$. Notice that in equation (5.27), the dependence of $\mathbf{L}$ on β is through equations (5.25) and (5.26).

For logistic regression, that is, when the logistic link in equation (1.3) is used, equation (5.27) simplifies to

$$\mathbf{L}(\beta) = \sum_{i \in \mathcal{F}} (y_i - p_i)\mathbf{X}_{i+}.$$

Of course $\mathbf{L}$ and *a fortiori* $\mathbf{B}$ are unknown. They are considered finite population parameters to be estimated from the sample. Using equations (5.22) and (5.23), for any β, $\mathbf{L}(\beta)$ can be estimated from the sample by

$$\begin{aligned}
\widehat{\mathbf{L}}(\beta) &= \sum_h \overline{\mathbf{z}}_h = \sum_h n_h^{-1} \sum_{r \in \mathcal{S}_h} \mathbf{z}_{hr} \\
\mathbf{z}_{hr} &= n_h \sum_k w_{hr} f_{hrk} \frac{y_{hrk} - p_{hrk}}{p_{hrk}(1-p_{hrk})} g^{'}(\eta_{hrk})\mathbf{X}_{hrk+} \\
\widehat{Var}(\widehat{\mathbf{L}}(\beta)) &= \sum_h n_h^{-1} \frac{\sum_{r \in \mathcal{S}_h} (\mathbf{z}_{hr} - \overline{\mathbf{z}}_h)(\mathbf{z}_{hr} - \overline{\mathbf{z}}_h)^T}{n_h - 1}.
\end{aligned} \tag{5.28}$$

The sample estimate $\widehat{\beta}$ satisfies

$$\widehat{\mathbf{L}}(\widehat{\beta}) = \mathbf{0}. \tag{5.29}$$

Using a Taylor-series linear approximation, $\widehat{\mathbf{L}}(\mathbf{B}) \approx -\frac{\partial \widehat{\mathbf{L}}}{\partial \beta}(\mathbf{B}) \cdot (\widehat{\beta} - \mathbf{B})$ where $\frac{\partial \widehat{\mathbf{L}}}{\partial \beta}(\mathbf{B})$ is a $(P+1) \times (P+1)$ matrix of partial derivatives evaluated at $\mathbf{B}$. This leads to the *linearization* estimator:

$$\widehat{Var}(\widehat{\beta}) = \left[\frac{\partial \widehat{\mathbf{L}}}{\partial \beta}(\widehat{\beta})\right]^{-1} \cdot \widehat{Var}(\widehat{\mathbf{L}}(\widehat{\beta})) \cdot \left[\frac{\partial \widehat{\mathbf{L}}}{\partial \beta}(\widehat{\beta})\right]^{-1^T} \tag{5.30}$$

where $\widehat{Var}(\widehat{\mathbf{L}}(\widehat{\beta}))$ is calculated by substituting $\widehat{\beta}$ for β in (5.28).

Notice that (5.30) estimates the variability, due to the sampling, of $\widehat{\beta}$ from the finite population parameter $\mathbf{B}$. We might be interested in the variability of $\widehat{\beta}$ as an estimate of the superpopulation parameter β in (5.26). This variability is due to both the sampling

of the finite population and the variability of the finite population as a realization of the superpopulation. Letting d (for design) and m (for model) denote expection under the sampling and superpopulation respectively,

$$\begin{aligned} Var_{m,d}(\widehat{\beta}) &= E_m(Var_d(\widehat{\beta})) + Var_m(E_d(\widehat{\beta})) \\ &\approx E_m(Var_d(\widehat{\beta})) + Var_m(\mathbf{B}) \\ &= E_m(Var_d(\widehat{\beta})) + \left[E_m(-\frac{\partial \mathbf{L}}{\partial \beta}(\beta))\right]^{-1}. \end{aligned} \tag{5.31}$$

The last equality in equation (5.31) is obtained by recollecting that $\mathbf{B}$ is the MLE of β in the superpopulation model and hence its variance, under the superpopulation model, is asymptotically given by the inverse of the information matrix (see for example Silvey [7]). Equation (5.31) leads to the estimate

$$\widehat{Var}_{m,d}(\widehat{\beta}) = \widehat{Var}(\widehat{\beta}) + \left[-\frac{\partial \widehat{\mathbf{L}}}{\partial \beta}(\widehat{\beta})\right]^{-1}. \tag{5.32}$$

We shall call this the *total-variance* estimator and $\widehat{Var}(\beta)$ the *design-based* variance estimator. Notice that the first term of (5.32) is of order n^{-1} whereas the second term is of order N^{-1}. Thus we expect the total variance to be only slightly larger than the design based variance.

Some would argue that $\widehat{Var}(\widehat{\beta})$ is already a total-variance estimator because the AFS sample design is closer to stratified random cluster sampling *without* replacement (see Kott [4]). Using the with-replacement variance formula in equation (5.28) when one has a without-replacement design implicitly assumes a model where the stratum population sizes are infinitely large. We, however, have found that $\widehat{Var}_{m,d}(\widehat{\beta})$ in equation (5.32) and the difference between $\widehat{Var}_{m,d}(\widehat{\beta})$ and $\widehat{Var}(\widehat{\beta})$ have, admittedly *ad hoc*, uses.

5.2 The modified stepwise regression methodology–main effects

The modified stepwise regression methodology incorporated in this study was designed to be conservative in the number of terms incorporated into the model. As illustrated by equation (1.1), when too many terms are incorporated into the model, the model yields nonsensical results.

In addition, and somewhat more subtly, the model became numerically unstable. This became especially apparent when models with interactions were fit. As discussed at the end of Section 1.1, if a fully saturated model were fit, that is, all possible interactions were potentially included, then the number of available parameters would far exceed the number of farms in the AFS. Thus all $\widehat{p}_i$ would be approximately 0 or 1, depending upon whether or not the i-th farm were on this list or not. What is observed is that when the model becomes too big, the fitted coefficients become quite large: that is $|\widehat{\beta}_j| > 10$, even for an *indicator* variable X_{+j} that only takes on the values 0 or 1. Notice that such a variable can cause a change in $\widehat{\eta}_i$ (see equation (1.16)) exceeding 10. Referring to Figure 1, such a change can change a fitted probability from essentially 0 to essentially 1. The model is overfitting the data in a more subtle version of the "solitary nursery in Stratum 32" problem discussed in Section 2.1.

We do not believe that any of the variables are that controlling, and models containing such controlling variables are useless for our purposes. This apparent misbehavior in parameter estimates was observed when only approximately 20 (out of more than 400 possible)

Table 8: Start of Modified Stepwise Regression Algorithm
(California data)

reduced model	full model	df	Wald χ^2	p-value
model0	model0.ftype	28	183.209279	0.000000e+000
model0	model0.ptfarm	1	2.618003	1.056573e-001
model0	model0.sales	10	118.661758	0.000000e+000
model0	model0.oper	6	71.652343	1.872946e-013
model0	model0.horse	5	43.247320	3.292195e-008
model0	model0.land	3	12.158624	6.859071e-003
model0	model0.strat	8	105.576634	0.000000e+000

state*sales interaction terms were added to the main effects 48-state models of Subsection 4.1. Not only were some of the estimated interaction terms unacceptably large, but the estimated coefficients for the main effects for the indicator variables of three states (Rhode Island, Indiana, and Nebraska) were sent out of the range $(-10, 10)$. Thus, caution in adding terms to the model is amply justified in our view.

Suppose the $P+1$ coefficients β are divided into two groups: $\beta = (\gamma_0, \gamma_1)$, where γ_0 has $Q+1$ coefficients and includes the intercept and γ_1 has $P-Q$ coefficents. To test $H_0 : \gamma_1 = \mathbf{0}$, we will use a Wald statistic,

$$\chi^2 = \widehat{\gamma}_1^T [\widehat{Var}(\widehat{\gamma}_1)]^{-1} \widehat{\gamma}_1, \tag{5.33}$$

where $\widehat{\beta} = (\widehat{\gamma}_0, \widehat{\gamma}_1)$ is defined by (5.29) and $\widehat{Var}(\widehat{\gamma}_1)$ is the lower $(P-Q)\times(P-Q)$ submatrix of $\widehat{Var}(\beta)$ as defined by (5.30). We shall often refer to this as a test of the "reduced model", defined by the terms corresponding to γ_0, versus the "full model", defined by the terms corresponding to β. The asymptotic critical point for (5.33) is a χ^2 distribution with $P-Q$ degrees of freedom.

Starting with a reduced model of the intercept only ("model 0" in what follows), each of the groups was tested to see whether it represented a significant improvement over the reduced model. The most significant group was selected first. For example, Table 8 shows the results for the California data set. The most significant groups were farm type, sales, and AFS stratum. For reasons described in Section 2, the stratum group was deferred. The sales group was chosen first, in preference to the farm-type group, because it had the larger χ^2 to *df* ratio.

Standard stepwise logistic regression (except that design-based variance estimates were used) with $\alpha = .05$ was then performed to determine which variables in the sales group should be added. This resulted in a model ("model1" in what follows) with

intercept + sales5K + sales50K + sales1000K.

Then starting with model1, each of the remaining groups was tested to see if it represented a significant improvement. The results are shown in Table 9.

The next group to visit should be the farm-type group. Notice also that the horse and land groups although highly significant in Table 8 were no longer significant in Table 9 ($\alpha = .05$). This means that the explanatory value of these two groups in modeling NML were already accounted for by the sales group. For example, although large-acreage farms

Table 9: Second Step of Modified Stepwise Regression Algorithm
(California data)

reduced model	full model	df	Wald χ^2	p-value
model1	model1.ftype	28	489.8464893	0.000000e+000
model1	model1.ptfarm	1	0.1522516	6.963929e-001
model1	model1.oper	6	48.7612409	8.322987e-009
model1	model1.horse	5	9.7614724	8.228292e-002
model1	model1.land	3	2.4360261	4.869636e-001
model1	model1.strat	8	91.6896310	2.220446e-016

are probably more likely to be on the list, large-acreage farms tend to have larger sales, and the model already predicts that farms with larger sales are more likely on the list.

Stepwise regression was then performed with the variables in the farm type group. During this process, the variables in model1 (*intercept + sales5K + sales50K + sales1000K*) were kept in the model. After the completion of the stepwise regression algorithm on the farm-type group, however, the three variables in the sales group were checked to see whether they remained significant. The resulting model, denoted by 'model2', is

intercept + sales5K + sales50K + sales1000K + CHRS + CCENFRUT + CCENCOTT + CCENSHEP + CCENAQUA.

This procedure was iterated until no groups represented statistically significant improvement, arriving at the main effects models given in Tables 4, 5, and 6.

5.3 The modified stepwise regression methodology–interactions

Because the 48-state AFS has 46,000 data points (farms), after fitting the main effects, it was decided to try to fit models with (two way) interactions. Thus, groups such as state*sales were considered for addition. Notice that these groups can be quite large; for example state*sales has 480 variables.

Several problems were noticed in applying the modified stepwise regression procedure to groups defined by interactions. In addition to the hidden-small-cell problems noted at the beginning of the previous subsection, the matrices $\widehat{Var}(\widehat{\gamma}_1)$ in the Wald statistics (5.33) were often numerically ill conditioned[11] These problems lead to values of the Wald statistic that were so huge that their p-values were meaningless. There were often large discrepancies

[11] *Ill conditioned* matrices are nonsingular matrices that are close to being singular. This leads to large round-off errors in the calculation of their inverses. For symmetric positive definite matrices, conditioning is often measured by ratio of the largest to the smallest eigenvalue; if the ratio is too big (say around 10^8), the matrix is ill conditioned. For our purposes, write $\widehat{Var}(\widehat{\beta}) = \sum_{m=1}^{m=P+1} \lambda_m \mathbf{e}_m \mathbf{e}_m^T$ be the eigendecomposition of $\widehat{Var}(\widehat{\beta})$, that is the decomposition of $\widehat{Var}(\widehat{\beta})$ into its principal components, where $\lambda_1 > \lambda_2 > \cdots > \lambda_{P+1} > 0$. Then the Wald statistic

$$\widehat{\beta}^T \left[\widehat{Var}(\widehat{\beta})\right]^{-1} \widehat{\beta} = \sum_m \lambda_m^{-1} (e_m^T \widehat{\beta})^2.$$

When λ_{P+1} is very small compared to λ_1, small sampling and roundoff variations in $e_{P+1}^T \widehat{\beta}$ lead to gigantic changes in the Wald statistic.

between the design-based-variance estimates (5.30) and their total-variance analogues (5.32) indicating a probable break down of the asymptotics. Even when extremely conservative approaches were used, such as Bonferroni corrections to a base signicance level of .01, too many interaction terms became significant, and the models became numerically unstable.

After much trial and error, the following *ad hoc* procedure was applied to each group of interactions. For the purposes of illustration we discuss the state*sales group. Notice that there are 48 state variables and 10 sales variables, which, in what follows, we denote by $state_r, \quad r = 1, \cdots, 48$ and $sales_s, \quad s = 1, \cdots, 10$.

Step 1. 48 models, one for each state, were fit to the entire 48-state data set. The variables in the r-th model consisted of the current model + $state_r \cdot sales_s, \quad s = 1, \cdots, 10$. Each of these state models was considered a significant improvement only when the Wald χ^2 statistic (using the total covariance instead of the design based covariance) was significant at a level of $.01/50 = .0002$.

Step 2. When the r-th state model was judged significant in Step 1, the variables of the form $state_r \cdot sales_s, \quad s = 1, \cdots, 10$ were examined to see which satisfy the following two criteria:

- The corresponding coefficient was significant at a $\alpha = .01/10 = .001$ level, using total-variance estimates to calculate standard errors.
- The estimated standard error of the coefficient calculated using the total-variance formula was no greater than twice the estimated design-based standard error. This criterion, for reasons which at present are poorly understood, seemed to prevent numerical instability.

If several variable satisfied these two criteria the most significant among them was selected.

Step 3. The variables identified in Step 2, at most one for each significant state model, were simultaneously added to the model and the model refit.

Step 4. Using the model fit of Step 3, a state*sales variable (whether added at this stage or any previous stage) was marked for deletion when it satisfied the following two criteria:

- The corresponding coefficient was no longer significant at a $\alpha = .01$ level, using the total-variance formula to estimate standard errors.
- The standard error of the coefficient estimated using the total-variance formula was greater than twice the estimated design-based standard error.

The model was then refit.

Steps 1-4 were iteratively applied until no further states*sales variables entered or left the model.

This procedure was then applied to the other two-way interactions which include the state group as one factor. After these two-way interactions were fit, the model was checked to see whether main effects for any states needed to be added or whether other effects became superfluous and could be deleted. The significance level used continued to be .05 for main effects and .01 for interactions.

6 Concluding Remarks

6.1 Summary of results

In our studies, we have seen that the logistic model for predicting whether a farm is NML appears to work well (at least in California) and that popular alternatives (the probit *et al.*) perform in a roughly equivalent manner. If the model is to be used for weight calibration, then integerization is a larger source of model degradation than truncating the weights at a fixed upper limit. Moreover, truncating weight from a logistic model appears at least as effective as using the bounded analogue of the logistic link function in equation (1.19).

For the 2002 Census, NASS effectively calibrated with a truncated variant of the inverse-linear link function (the inverse-linear link is displayed in equation (1.18)). Although our investigations did not address the inverse linear link *per se* or its implicit estimation through calibration, the results on integerization and truncation are somewhat relevant. Truncation hardly mattered with a logistic link, although an inverse-linear link might not fare as well. Integerization, by contrast, weakened the logistic-model fit noticeably.

Logistic regression has potential for predicting where to find NML farms in future area-frame surveys as described in Subsection 1.2. Since NASS sample designs are independent across states, it is tempting to fit a separate logistic model for each state. We, however, prefer that a single 48-state model be used for this purpose because of its larger sample size.

Based on our 48-state analysis, best predictors of whether a 2002 farm was NML, all other things being equal, were

low sales (the lower the sales the more likely a farm being NML);
a black, Asian, Hispanic, or woman principal operator (black especially);
primary sales from Christmas trees or equine;
the presence of horses for personal or "other" use;
the the presence of nursery products; and
being in AFS Stratum 30 or higher.

That is to say, a farm with a black principal operator, say, was more likely to be NML in 2002 than other farms in the same sales class and AFS stratum.

6.2 *Caveats* and warnings

It is important to realize that a variable's failing to appear in the model, does not mean that the variable is not predictive of NML status. The variables exhibit a great many relationships among themselves and it is possible that a highly predictive variable is related to a different variable which does appear in the model.

Secondly, the authors believe that the hidden-small-cell problem discussed in Subsections 2.1 and 5.2 cannot be dismissed. As a result, the 48-state models are preferable to the individual state models.

Finally, model-fitting procedures are highly subjective. Different procedures often yield different models. The authors believe that the procedures describe herein are reasonable, consistent with existing statistical procedures for model fitting in the context of linear regression, and computationally feasible. This does not mean that there do not exist other procedures and other models.

6.3 Suggested methodology questions for future research

The hidden-small-cell problem and the related issue of numerical instability in the parameter estimates need further exploration. The problem even arises in large data sets (46,000 observations) with a moderate number of variables (about 50).

Empirically, this study indicates that the problem often manifests itself in a discrepancy between what we called (somewhat dubiously) the design-based variance estimator and the sum of that estimator and an estimator for the model variance of the finite-population logistic-regression coefficient ($Var_m(\mathbf{B})$ in equation (5.31)). As a result, this study shows the need for research into why this appears to be so, as well as the consideration of alternative model-based variance estimators.

Another side of the hidden-small-cell problem is that some number of observations can be highly influential on some linear combination of the parameter estimates. Development of influence function methodology for the sample survey-context to detect this phenomenom is a promising line of further inquiry.

Appropriate remedial analyses should be developed for surveys that exhibit hidden-small-cell problems. In a stratified cluster sample, Zaslavsky [8] used influence-function technology to detect overly influential PSU's and then reweighting to lessen the influence of such PSU's. For the NASS AFS, however, influential observations are at a sub PSU (tract) level and hence Zaslavsky's influence function definition does not directly apply. Furthermore, influential observations outside the survey-sampling context are often handled by using estimating functions (such as L_1 sum of distances) in place of least squares (which corresponds to simply taking means). This suggests that alternative approaches to reweighting might be fruitful for handling overly influential observations. More research can be productively done in this direction.

It should be noted that in the NASS AFS often small farms have large weights relative to large farms. This is because the total weight in the NASS AFS survey has two components: a sampling weight and a tract acreage to farm acreage ratio. Often, for small farms, the tract/farm acreage ratio is close to 1. But for large farms, it is often quite small (less than .01). Thus it is entirely possible that for some parameters (or some hidden linear combination of parameters), the small farms are excessively influential relative to the large farms, and the development of techniques to uncover such situations would seem to be compelling.

Finally, this information can be used to to design future surveys that avoid the hidden-small-cell problem where possible.

7 Acknowledgements

The author wishes to thank Herb Eldridge for preparing the data file Bill Wigton for his many insights into the Area Frame Survey.

8 Appendix: computer code

8.1 SAS code

8.2 Splus code

[Available upon request.]

References

[1] Folsom, R. E. and A. C. Singh (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *ASA Proc. Surv. Res. Meth. Sec.*, 598-602.

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

[3] Kott, P. S. (2004). Using calibration weighting to adjust for nonresponse and coverage errors. Pesented at the Mini Meeting on Current Trends in Survey Sampling and Official Statistics, Calcutta, India. paper

[4] Kott, P. S. (1988). *Estimating Variances for the June Enumerative Survey.* US Department of Agriculture, National Agricultural Statistics Service, SRB Staff Report Number SRB-88-06.

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

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

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

[8] Zaslavsky, A. M., N. Schenker, and T. R. Belin (2001). Downweighting influential clusters in surveys: application to the 1990 post enumeration survey. *J. Amer. Statist. Soc.* **96**, 858-869.