Adjusting the June Area Survey for Non-response and Misclassification

Kenneth K. Lopiano* Denise Abreu† Pam Arroway‡ Andrea C. Lamas § Linda J. Young¶

Abstract
In years ending in 2 and 7, the National Agricultural Statistics Service (NASS) conducts the Census of Agriculture. In addition, NASS conducts an annual area-frame-based survey called the June Area Survey (JAS). Both the Census of Agriculture and the JAS provide an estimate of the number of farms in the United States. In 2007, the difference between the two estimates could not be attributed to the error associated with the respective estimates. In this paper, non-response and misclassification in the JAS are considered as possible factors leading to the discrepancy. How to account for misclassification and non-response in the JAS-based estimate is discussed.

Key Words: Non-response, June Area Survey, Misclassification

1. Introduction

Each year, the National Agricultural Statistics Service conducts the June Area Survey. In addition, NASS conducts the Census of Agriculture in years ending in 2 and 7. In 2007, both the JAS and the Census of Agriculture provided an estimate of the number of farms in the United States. The difference between the estimates could not be attributed to the error associated with the respective estimates. Two factors associated with the JAS that have been explored as possible contributors to the discrepancy are non-response and misclassification. In this paper we present an approach that can be used to adjust the JAS for non-response and misclassification. In section 2, a description of the estimation of agricultural activities is presented, followed by an explanation of the concerns associated with estimation. In section 3, an explanation of how non-response can be used instead of estimation is provided. In sections 4 and 5, a review of the procedure used to identify misclassification and how misclassification and non-response can be adjusted for simultaneously is presented. Finally, in the last section, a discussion regarding future directions is provided.

2. Estimation

One factor associated with the JAS that contributes to the discrepancy between the two estimates arises from the estimation of agricultural activities for sampled tracts. During the sample selection process, tracts of land in the United States are selected to be surveyed for agricultural activity. When a tract operator is either inacccesible for a JAS interview or refuses to participate in the JAS, enumerators are instructed to estimate the tract-level agricultural items. As a result, farm-level items are left to be imputed. The farm-level items are imputed using other sources (NASS surveys, previous JASs, Farm Service Agency information, etc.) or using imputation methodologies.

One farm-level item that is imputed is the tract-to-farm ratio. When calculating the total number of farms, the tract-to-farm ratio (the tract acreage divided by the total farm acreage) is used to represent the

*Department of Statistics, University of Florida, Gainesville, FL 32611
†National Agricultural Statistics Service, USDA, 3251 Old Lee Hwy, Fairfax VA 22030
‡Department of Statistics, North Carolina State University, Raleigh, NC 27695
§National Agricultural Statistics Service, USDA, 3251 Old Lee Hwy, Fairfax VA 22030
¶Department of Statistics, University of Florida, Gainesville, FL 32611

proportion of a farm present in a tract. When the agricultural activity in a tract is estimated, the tract-to-farm ratio is imputed using either previously reported/administrative data or a median imputation approach. Although median imputation was a common solution when this problem was first addressed, more recent research has illustrated its limitations. Therefore, estimation of data for non-response tracts was identified as a potential area for improvement in the process of estimating the number of farms in the United States.

In 2009, NASS conducted the Farm Numbers Research Project (FNRP). In this study, 595 estimated tracts' farm-level items from the JAS and the FNRP survey were compared. A high level of discordance was observed for a number of variables, including tract-to-farm ratio and total farm acreage. Although the quality of the imputed data for total farm acreage is likely related to the method of imputation, the current NASS procedures do not capture information on the source of the imputed value. Knowing the source of the information used to impute would provide insight regarding the quality of the other farm-level items that were also imputed. Because the quality of imputed values for estimated tracts cannot currently be determined, an intermediate solution is to treat each estimated tract as a unit non-respondent. Then, the JAS-based estimate of the number of farms can be adjusted using unit non-response methodologies. Such an approach, although statistically viable, is not able to fully utilize the information collected from estimated tracts. Because a portion of the imputed values are based on reliable sources, amending the JAS is necessary to ensure the retainment of quality estimated tracts and a proposal to make such revisions in the JAS has been submitted. In the future, tracts with quality information can be treated as respondents and those remaining will be treated as unit non-respondents. Nonetheless, the option of using a non-response methodology to adjust the June Area Survey is considered here because even after obtaining the information regarding quality estimated tracts, non-response will still be present.

3. Non-response

3.1 Non-response Model

The current estimate for the number of farms based on the JAS can be simplified to the following expression,

$$T = \sum_{i \in R} \pi_i^{-1} y_i t_i,$$

where R denotes the set of respondents, π_i denotes the inclusion probability of respondent i, $y_i = 1$ if the tract contains a farm and is 0 otherwise, and t_i = tract-to-farm ratio. If ϕ_i denotes the probability of response for unit i, then the non-response weighted estimate for the total number of farms would be

$$T_{NR} = \sum_{i \in R} \pi_i^{-1} \phi_i^{-1} y_i t_i.$$

In practice, ϕ_i is unknown and must be estimated. ϕ_i can be estimated in several ways.

Although sampling weights have often been incorporated in non-response methodologies (Platek and Gray, 1983), Little and Vartivarian (2003) show that "weighting response rates by sampling weights to adjust for design variables is either incorrect or unnecessary." (pp. 1589) The authors instead suggest that the correct approach is to model non-response as a function of covariates and design variables. Given the model, the response weight is the inverse of the estimated probability from this model.

3.2 Estimating ϕ: Logistic Regression

Building on the recommendations described in the previous section, a logistic regression model was developed to estimate the probability of responding to the JAS. During the JAS, tract-level items are recorded for

both respondents and non-respondents. For tract-level items, a simple binary indicator of the presence or absence was used as a covariate. For example, if an enumerator observes corn in a tract then the corn indicator for that tract was 1. In addition, state and land-use strata were used as covariates. State and land-use strata are common to both respondents and non-respondents, and are design variables used in the sample selection procedure. The land-use strata variable takes one of five values: 50% cultivated, 15 to 50% cultivated, agricultural urban/commercial, less than 15% agricultural or non-agricultural. The final logistic regression model can be expressed as follows; for a given tract,

$$\begin{aligned} Z_i &\sim Bernoulli(\phi_i) \\ \text{logit}(\phi_i) &= X_i\boldsymbol{\beta} \end{aligned}$$

where Z_i is 1 if the operator of the i^{th} tract responded and is 0 otherwise, X_i is the vector of covariates for the i^{th} tract and $\boldsymbol{\beta}$ is a vector unknown regression coefficients. An additive model with only main effects was assumed and backwards elimination was used to remove insignificant covariates.

Based on the logistic regression model, $\hat{\phi}$ is estimated for each respondent and incorporated in the non-response model. That is, the estimated non-response adjusted estimate is given by

$$\hat{T}_{NR} = \sum_{i \in R} \pi_i^{-1} \hat{\phi}_i^{-1} y_i t_i,$$

where $\hat{\phi}_i$ is the estimated response probability for tract i.

3.3 Variance Estimation

The additional uncertainty due to non-response must be accounted for when estimating the variance of the estimate. Moreover, the error in estimating the response propensity must be accounted for in variance calculations. The methodology of Kim and Kim (2007) provides a framework for estimating the variance of design based estimates adjusted for non-response.

Recall, the usual estimate of the number of farms is given by,

$$T = \sum_{i \in R} \pi_i^{-1} y_i t_i.$$

The non-response adjusted estimate of the number of farms is given by,

$$T_{NR} = \sum_{i \in R} \pi_i^{-1} \phi_i^{-1} y_i t_i.$$

Finally, the estimated non-response adjusted estimate is given by,

$$\hat{T}_{NR} = \sum_{i \in R} \pi_i^{-1} \hat{\phi}_i^{-1} y_i t_i,$$

where $\hat{\phi}_i$ is the estimated response probability for tract i. Kim and Kim (2007) show that under certain

assumptions, the variance of this estimate is estimated using the following formula,

$$\widehat{V(\hat{T}_{NR})} = \sum_{i \in R} \frac{1-\pi_i}{\pi_i^2} \hat{\phi}_i^{-1} (y_i t_i)^2 + \sum_{i \neq j \in R} \sum_{j \in R} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \hat{\phi}_i^{-1} \hat{\phi}_j^{-1} (y_i t_i y_j t_j) + \sum_{i \in R} \pi_i^{-2} \frac{(1-\hat{\phi}_i)}{\hat{\phi}_i^2} (y_i t_i - \pi_i \hat{\phi}_i \hat{\mathbf{h}}_i^T \hat{\gamma})^2,$$

where

$$\hat{\mathbf{h}}_i = \partial \text{logit}(\phi_i) / \partial \boldsymbol{\beta}(\hat{\boldsymbol{\beta}}) = X_i'$$

and

$$\hat{\gamma} = \left(\sum_{i \in R} (1-\hat{\phi}_i) \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^T \right)^{-1} \sum_{i \in R} \pi_i^{-1} \hat{\phi}_i^{-1} \hat{\mathbf{h}}_i y_i t_i.$$

4. Misclassification

Misclassification is another factor that contributed to the discrepancy observed in 2007. Misclassification occurs when a tract is incorrectly identified as non-farm when there is actually some portion of a farming operation inside the tract. Because the census data are responses to a list-based survey whose mailing list is created and maintained completely independently of the JAS area frame, the census data can be used to assess misclassification in the JAS. To do this, an attempt was made to link or match each 2007 JAS tract to its 2007 Census record. Disagreement in farm status between the JAS and census occurs when (1) tracts identified as non-farms in the JAS are subsequently identified as farms in the census or (2) tracts identified as farms in the JAS are subsequently identified as non-farms in the census. Here it is assumed that a tract that is identified as a farm in either the JAS or the census is a farm. Although misclassification of both types (1) and (2) are of interest, only the tracts identified as non-farms in the JAS are discussed in this paper. As a result of linking the JAS and the Census respondents list, the correct farm status and tract-to-farm ratios are obtained from a set of matched records. In years when a census is not conducted, a different methodology must be implemented. For discussion of one such methodology see Young *et al.* (2010) and Lamas *et al.* (2010).

5. Combining Non-response and Misclassification

Let U denote the set of respondents to the JAS after the records were updated using the information obtained from matching. U contains records that were misclassified as non-farms in the JAS but were identified to be part of a farming operation in the census. Note: records that were identified as non-respondents were not considered in the matching process because they are adjusted for using the non-response weights.

The response propensity for each tract in U is estimated under the modeling framework as described in the previous section. With $\hat{\phi}_i$ for all records in U, the non-response, misclassification adjusted estimate for the number of farms in the United States is given by

$$\hat{T}_{NR.M} = \sum_{i \in U} \pi_i^{-1} \hat{\phi}_i^{-1} y_i t_i.$$

Note, this estimate still potentially represents an undercount, as not all JAS records were matched to the census respondents list. It is *potentially* an undercount because it is possible that some of the JAS non-farm records that did not match to a census record could be farms, leading to an undercount.

The variance of this estimate is calculated using

$$\begin{aligned}\widehat{V(\hat{T}_{NR.M})} = &\sum_{i\in U}\frac{1-\pi_i}{\pi_i^2}\hat{\phi}_i^{-1}(y_i t_i)^2 \\ &+\sum_{i\neq j\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\hat{\phi}_i^{-1}\hat{\phi}_j^{-1}(y_i t_i y_j t_j) \\ &+\sum_{i\in U}\pi_i^{-2}\frac{(1-\hat{\phi}_i)}{\hat{\phi}_i^2}(y_i t_i - \pi_i\hat{\phi}_i\hat{\mathbf{h}}_i^{`}\hat{\gamma})^2.\end{aligned}$$

6. Conclusion and Future Work

In this paper, we identified estimation of data for individual tracts as an area of concern in the JAS. Based on our current state of knowledge, non-response methodologies have been used whenever agricultural activities were estimated. A framework for adjusting the JAS for non-response and misclassification was developed by assuming that each tract has a certain probability of responding to the survey. The probabilities were estimated using logistic regression and estimates with appropriate measures of uncertainty were obtained. When implementing the methodologies presented for the 2007 JAS, the non-response/misclassification adjusted estimate was within error of the 2007 Census of Agriculture estimate. It is worth noting, however, that the work presented here is a report of preliminary results of an ongoing research team. More sophisticated model building procedures will be considered when completing the work.

Future work adjusting the JAS for non-response needs to be developed. In subsequent years, the source of estimated farm level data will be obtained. As a result, tracts will fall into one of three categories: respondents, quality estimated tracts, and non-respondents. The quality estimated tracts will either have a quality estimated tract-to-farm ratio or enough information associated with the tract to implement a more sophisticated imputation methodology. Because non-response and misclassification will still be present in future JASs, the methodology presented in this paper provides a starting point for the development of a revised JAS-based estimate of the number of farms in the United States.

7. Acknowledgements

The authors of this paper are all members of a team brought together under a cooperative agreement between the National Institute of Statistical Sciences and USDA's National Agricultural Statistics Service.

REFERENCES

Abreu, Denise A., Pam Arroway, Andrea C. Lamas, Kenneth K. Lopiano, and Linda J. Young. 2010. Using the Census of Agriculture list frame to assess misclassification in the June Area Survey. Proceedings of the Joint Statistical Meetings.

Kim, JK and Kim, JJ . Non-response weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*. Vol. 35, No. 4, 2007, Pages 501−514

Lamas, Andrea C., Denise Abreu, Pam Arroway, Andrea C. Lamas, Kenneth K. Lopiano, and Linda J. Young. 2010. Modeling misclassification in the June Area Survey. Proceedings of the Joint Statistical Meetings.

Little, R.J.A. and Vartivarian, S. On weighting the rates in non-response weights. *Statistics in Medicine*. 2003; 22:1589−1599

Platek R, Gray GB. Imputation methodology. In *Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies*, Madow WG, Olkin I, Rubin DB (eds). Academic Press: New York, 1983; 255294

Young, Linda J., Denise Abreu, Pam Arroway, Andrea C. Lamas, and Kenneth K. Lopiano. 2010. Precise Estimates of the Number of Farms in the United States. Proceedings of the Joint Statistical Meetings.