ABSTRACT

Title of Document: FAY-HERRIOT SMALL AREA ESTIMATION IN THE SURVEY OF BUSINESS OWNERS.

Aneesah N. Williams,
Master of Arts, 2007

Directed By: Professor Eric V. Slud
Statistics Program
Department of Mathematics

This paper will study the use of the Fay-Herriot small area estimation model on the Survey of Business Owners and Self-Employed Persons (SBO) 2002 data. Small area estimation continues to be an important topic as the demand for reliable small area statistics continues to grow. Because direct estimates can yield large standard errors due to small sample sizes, the need for small areas to borrow strength from related areas is present. The 2002 SBO has a state level design, which may contain several counties per state with small sample sizes. This paper investigates the plausibility and usability of the Fay-Herriot estimators at the county level for predicting Black ownership of businesses. These mixed-effect model predictors will be compared to linear fixed-effect models. This research will also investigate the parsimony of the two types of estimators.

FAY-HERRIOT SMALL AREA ESTIMATION IN THE SURVEY OF BUSINESS OWNERS

By

Aneesah N. Williams

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Arts
2007

Advisory Committee:
Professor Eric V. Slud, Chair
Professor Paul J. Smith
Professor Ben N. Kedem

Dedication

To the two Maurices: my strength and my inspiration.

Acknowledgements

I want to give a very special thanks to the SBO staff in the Company Statistics Division of the U.S. Census Bureau. Without your assistance and advice, this project would not have been possible. Thank you for being a great reference and for contributing your resources to my research efforts.

Table of Contents

List of Tables

List of Figures

Chapter 1: Overview

1.1 Plan of the Study

The study was performed to examine the use of Fay-Herriot small area estimation models on the Survey of Business Owners and Self Employed Persons. The most recent survey data is from 2002, from which estimates of the race, gender, and Hispanic or Latino origin of the nation's business owners were produced. This paper describes the methodology for modeling and predicting county-level proportions of Black-ownership of businesses. It will present an evaluation of those predicted estimators as calculated using a linear, fixed-effect model and a Fay-Herriot small area model.

Chapter 1 begins by detailing the general methodology behind small area estimation. Typically, an area is regarded as large if the sample is big enough to yield direct estimates of adequate precision. Otherwise, an area is regarded as small (Rao, 2003). Often, there are many areas of interest (such as counties) that have a zero sample size. In making estimates for small areas, it is sometimes necessary to "borrow strength" by using values of the variable of interest from related areas, thus increasing the "effective" sample size (Rao, 2003).

These techniques were applied to the 2002 Survey of Business Owners and Self Employed Persons. Chapter 2 explains the scope of the survey and the sources of the data. The availability of good auxiliary data and the determination of a

suitable model are crucial to the formation of indirect estimators (Rao, 2003). A thorough description of the auxiliary data used in this study is given as well.

Chapter 3 describes the procedure for obtaining a linear, fixed-effect model for producing estimates of Black-owned businesses. The process involved many steps using automatic model selection techniques to obtain an adequate model. Thorough research showed that modeling at the county level, rather than the unit level, proved to be more predictive. Also, subdividing the data based on employer status and the presence of auxiliary data had a profound effect on the success and value of the fitted model. It was essential to fit a highly predictive, parsimonious model. In turn, the same predictors were to be used in the small area model prediction.

Chapter 4 describes the process of fitting the Fay-Herriot small area model and presents the results of the small area estimation. In the absence of any type of external validation of the predictors, lessons learned include under what circumstances and by how much the small area predictions alter the direct estimates. A comparison of the linear model's estimators to the Fay-Herriot model's estimators gave an indication of the utility of performing such work.

Although much is learned about the usefulness of small area estimation in the Survey of Business Owners and Self Employed Persons, there is still much work to be done. This study suggests such research could benefit estimation procedures in future SBO surveys.

Chapter 2: Small Area Estimation

2.1 Introduction

Small area statistics involves a wide variety of methods for drawing inferences about geographical or other subdomains of a survey. Often, national surveys are designed to ensure that inferences can be made about the main domain, and possibly a few subdomains such as states and counties. Such inferences would be design based using only the observed values of the variables. The usual direct estimators for subdomains, therefore, are likely to give unacceptably large standard errors because of the small samples sizes in those areas (Ghosh & Rao, 1994).

Most often, an overall sample size well above what can be afforded is required in order to make inferences about the lower level domains. Small area estimation attempts to solve this problem by using information from outside the subdomain, from values of other variables in that subdomain, and from information obtained outside the survey (Longford, 2005).

The main idea of small area techniques is exploiting similarity. To that end, the fact that the subdomain level means are similar to each other is exploited when estimating the district-level population mean of a recorded variable. The first step in the process is to determine how similar the districts are (Longford, 2005). Related to exploiting the similarity is borrowing strength across the subdomains.

This is done using a more traditional model-based approach that specifies a hierarchical model. The modeling approach is quite powerful. However, as Longford (2005) points out, the results it yields are heavily reliant on the validity of the model.

2.2 *Why Small Area Estimation is Needed*

The sampling design of a typical national survey seeks to ensure that inferences can be made with sufficient precision for the nation and possibly for the country's regions, or even states. Prescribing the sample sizes for each of several hundred small areas, such as counties or cities, is rarely feasible. In order to make the desired inferences within the small area, a subsample size much greater than what can be afforded would be necessary. This problem is addressed by drawing on auxiliary information from other areas, other variables, or from outside the survey (Longford, 2005).

Two types of small area models are often used, although this study focuses solely on the first of the models. In the first, area-specific auxiliary data are available and the parameters of interest are assumed to be related to the auxiliary data (Ghosh & Rao, 1994). In this study, the area-specific auxiliary data are aggregated to the county level, given as a county-level proportion, and then transformed with the logit function. The assumption is made that

$$\theta_i = x_i^T \beta + v_i z_i \qquad (2.1)$$

where the x_i are county-level auxiliary data,

the z_i's are known positive constants,

β is the vector of regression parameters, and

the v_i's are independent, identically distributed (i.i.d.) random variables with $E(v_i)=0$, $Var(v_i)=\sigma_v^2$.

In the second model, a nested error regression model, Ghosh & Rao (1994) explain that unit-specific auxiliary data are available for the population elements, and the variable of interest is assumed to be related to x_{ij} through a regression model given as

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij} \qquad (2.2)$$

where

$j=1, \ldots, N_i$,

$i=1, \ldots, m$,

$e_{ij} = \tilde{e}_{ij} k_{ij}$ and $E(\tilde{e}_{ij}) = 0$, $Var(\tilde{e}_{ij}) = \sigma^2$,

the k_{ij}'s are known constants, and

N_i is the number of elements in the i^{th} area.

The symbol *Y* denotes the vector representing the values of the variable of interest. The survey data on this variable is *y*, and $\hat{\theta} = \hat{\theta}(y)$ is an unbiased estimator of the parameter of interest. In many instances, $\hat{\theta}$ is the estimated population total or mean of the variable of interest. We assume $\hat{\theta}_i$ is unbiased for θ_i, the parameter for any subdomain *i* (Longford, 2005).

The mean squared error (MSE) of the estimator $\hat{\theta}$ of the target θ is defined as $MSE(\hat{\theta};\theta) = E\left[\left(\hat{\theta}-\theta\right)^2\right]$. The θ in MSE is vital because $\hat{\theta}$ may be used for estimating θ_i.

The MSE of the parameter, $\hat{\theta}_i$, is given as

$$M_{1i}\left(\sigma_v^2\right) = E\left[\hat{\theta}_i - \theta_i\right]^2 = g_{1i}\left(\sigma_v^2\right) + g_{2i}\left(\sigma_v^2\right) \tag{2.3}$$

where

$$g_{1i}\left(\sigma_v^2\right) = \sigma_v^2 z_i^2 \psi_i \left(\sigma_v^2 z_i^2 + \psi_i\right)^{-1} = \gamma_i \psi_i \text{ and}$$

$$g_{2i}\left(\sigma_v^2\right) = \left(1 - \gamma_i\right)^2 x_i^T \left[\sum_i x_i x_i^T / \left(\sigma_v^2 z_i^2 + \psi_i\right)\right]^{-1} x_i .$$

In this equation, γ_i measures the uncertainty in modeling the θ_i's and ψ_i represents the sampling variance.

The MSE is regarded as a measure of efficiency, so an estimator with a smaller MSE is preferred. However, MSE usually must be estimated, its value possibly depending on one or more parameters or the target itself (Longford, 2005). An approximately unbiased estimator of the mean squared error is

$$MSE\left(\hat{\sigma}_v^2\right) = g_{1i}\left(\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_v^2\right) \tag{2.4}$$

where

$$g_{3i}\left(\hat{\sigma}_v^2\right) = \psi_i^2 z_i^2 \left(\sigma_v^2 z_i^2 + \psi_i\right)^{-3} \overline{V}\left(\hat{\sigma}_v^2\right).$$

In this equation, $\overline{V}\left(\hat{\sigma}_v^2\right)$ is the asymptotic variance of $\hat{\sigma}_v^2$.

2.3 Fay-Herriot Model

The Fay-Herriot model (Fay and Herriot, 1979) for small area estimation uses a linear mixed-effect model of the form given in Equation (2.1). Typically, normality of the fixed and random effect terms is assumed. It is also customary to

assume the sampling variances are known. The area-specific auxiliary data, x_{ij}, are available for the population elements.

Most small area models are special cases of generalized linear models with both fixed and random effects. Small area parameters can be expressed as linear combinations of these effects (Ghosh & Rao, 1994). The best linear unbiased predictor (BLUP) estimators of these parameters minimize the MSEs and are not dependent on normality of the subdomain level effects and sampling errors.

The BLUP estimator of θ_i is simply a weighted average of the direct estimator $\hat{\theta}_i$ and the regression synthetic estimator $x_i^T \beta$. Therefore, the BLUP estimator takes into account the between area variation and the precision of the direct estimator (Ghosh & Rao, 1994). Specifically,

$$\tilde{\theta}_i = \gamma_i \hat{\theta}_i + (1-\gamma_i) x_i^T \tilde{\beta} \qquad (2.5)$$

where $\tilde{\beta} = \left[\sum_i x_i x_i^T / \left(\sigma_v^2 z_i^2 + \psi_i \right) \right]^{-1} \left[\sum_{i=1}^{m} x_i \hat{\theta}_i / \left(\sigma_v^2 z_i^2 + \psi_i \right) \right]$ is the best linear unbiased estimator (BLUE) of beta and

$\gamma_i = \sigma_v^2 z_i^2 / \left(\sigma_v^2 z_i^2 + \psi_i \right)$.

The BLUP estimator is dependent on the variance component σ_v^2. However, replacing σ_v^2 with an asymptotically consistent estimate $\hat{\sigma}_v^2$ yields a two-stage estimator (Ghosh & Rao, 1994). This estimator, $\hat{\theta}_i$, is called the empirical BLUP, or EBLUP. It will remain approximately unbiased provided that the distributions

of v_i and e_i are both symmetric, though not necessarily normal. Additionally, $\hat{\sigma}_v^2$ must be an even function of $\hat{\theta}_i$ and remain invariant when $\hat{\theta}_i$ is changed to $\hat{\theta}_i - x_i^T a$ for all a (Kackar & Harville, 1984).

Chapter 3: The Survey of Business Owners and Self-employed Persons

3.1 The Survey

The Survey of Business Owners and Self-employed Persons (SBO) is a consolidation of two prior surveys, the Survey of Minority-Owned Business Enterprises (SMOBE) and the Survey of Women-Owned Business Enterprises (SWOBE). It also includes questions from a survey discontinued in 1992 on Characteristics of Business Owners (CBO).

The SBO is part of the Economic Census, which is conducted every five years. The most recent year for which SBO data are available is 2002. SBO statistics describe the characteristics of U.S. businesses by gender, race, and Hispanic or Latino origin of the principal owners; by geographic area at the national, state, and sub-state regional levels; by two-digit industry sector based on the 2002 North American Industry Classification System (NAICS); by size of firm (employment size and receipt size); and by employment status (Census, 2007).

NAICS is an industry classification system that groups establishments into industries based on the activities in which they are primarily engaged. It is a comprehensive system covering the entire field of economic activities, producing

and nonproducing. There are 20 NAICS sectors in the United States (Executive Office of the President, 2002).

3.2 Sources of the Data

A random sample of businesses was selected from a list of all firms operating during 2002 with receipts of $1,000 or more. The SBO was conducted on a firm (company) basis rather than an establishment basis. A firm is a business consisting of one or more domestic establishments that the reporting firm specified as being under its ownership or control at the end of 2002. The universe of all firms was compiled from a combination of business tax returns and data collected on other economic census reports (Census, 2007). The Census Bureau obtained electronic files from the Internal Revenue Service (IRS) for all companies filing IRS form 1040, Schedule C (individual proprietorship or self-employed person); form 1065 (partnership); form 1120 (corporation); and form 941 (employer's quarterly federal tax return).

Firms in the following NAICS industries were considered out-of-scope to SBO and were therefore excluded from the sample:

- crop and animal production (NAICS 111, 112),
- domestically scheduled airlines (NAICS 481111),
- railroads (NAICS 482),
- U.S postal service (NAICS 491),
- mutual funds (NAICS 525) except real estate investment trusts (525930),
- religious grant operations (NAICS 813),

- private households and religious organizations (NAICS 814),
- public administration (NAICS 92).

SBO data on businesses included the number of firms, sales and receipts, annual payroll, and employment for firms in each ownership and geographic category. SBO statistics also identified family businesses, home-based businesses, types of customers and workers, sources and purposes of financing, and owner's age, education level, veteran status, and primary function(s) in the business.

3.3 Sampling Methodology

In designing the 2002 SBO sample, the following sources of information were used to estimate the probability that a business was minority- or woman-owned:

- administrative data from the Social Security Administration,
- lists of minority- and women-owned businesses published in syndicated magazines, located on the internet, or disseminated by trade or special interest groups,
- word strings in the company name indicating possible minority ownership (derived from 1997 survey responses),
- racial distributions for various state-industry classes (derived from 1997 survey responses) and racial distributions for various ZIP codes,
- gender, race, and Hispanic or Latino origin responses of a single-owner business to a previous SBO survey or to the 2000 Decennial Census.

These probabilities were then used to place each firm in the SBO universe in one of nine frames for sampling. The nine frames were American Indian, Asian, Black or African American, Hispanic, Non-Hispanic white male, Native Hawaiian and Other Pacific Islander, Other (a different race was supplied as a write-in to another source), Publicly-owned, and Women.

Each SBO company was placed into one of nine frames before sampling. To determine to which frame a company belonged, 12 predicted probabilities for race and 2 predicted probabilities for Hispanic or Latino origin were assigned to each record. These probabilities were estimated using a logistic regression that used administrative data, prior survey data, and consumer information (Galvin, 2006).

The SBO universe was stratified by state, industry, frame, and whether the company had paid employees in 2002. Large companies, including those operating in more than one state, were selected with certainty. These firms were selected based on volume of sales, payroll, or number of employees. All certainty cases had a selection probability and sampling weight of one. The certainty cutoffs for sales, payroll, and employees varied by sampling stratum and each stratum was sampled at varying rates, depending on the number of firms in a particular industry in a particular state. The remaining universe was subjected to stratified systematic random sampling.

3.4 Nonresponse

Approximately 81 percent of the 2.3 million businesses in the SBO sample responded to the survey. A firm was considered a respondent if it provided race,

gender, or Hispanic or Latino origin information for at least one owner, or if the firm was publicly held. Nonrespondents were first matched to the 1997 survey responses to obtain gender, Hispanic or Latino origin, and race data of the firms that were in both the 1997 and 2002 samples. Remaining nonrespondent data were imputed from donor respondents with similar characteristics of state, industry, employment status, size, and sampling frame. This nearest neighbor imputation method was used to impute gender, Hispanic or Latino origin, and race only (Williams, 2005). In this research, the inaccuracy and variability due to imputation was disregarded, as all data values were treated as though they were reported.

3.5 Tabulation

For SBO purposes, business ownership was defined as having 51 percent or more stock or equity in the business (Schlein, 2005). The various categories of businesses included gender: male, female, or equally male/female owned; ethnicity: Hispanic or Latino, or not Hispanic or Latino; and race: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander. Firms could be tabulated in more than one race category due to multiple race reporting.

For instance, a business with owner 1 reporting 70% ownership for a Non-Hispanic, Asian male and owner 2 reporting 30% ownership for a Hispanic, white female would be tabulated as a Non-Hispanic-owned business, a male-owned business, and an Asian-owned business. On the other hand, a business with

owner 1 reporting 50% ownership for a Hispanic, White/Black female, and owner 2 reporting 50% ownership for a Hispanic, American Indian/Black female would be tabulated as a Hispanic-owned business, a female-owned business, and a Black-owned business. A business with owner 1 reporting 100% ownership for a Non-Hispanic, Asian/Black male would be tabulated as a Non-Hispanic-owned business, a male-owned business, an Asian-owned business, and a Black-owned business.

3.6 Variance Estimation

Random groups were used to estimate the variances for the estimates produced in the 2002 SBO. Kish (1965) states that it is not necessary to assume independence between the selections that comprise a group. The variances for characteristics reported in the 2002 SBO were calculated using the following formula:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \left(\tfrac{1}{n}\right)\left(\tfrac{1}{n-1}\right)\left(\hat{y}_i - \hat{\bar{y}}\right)^2 \qquad (3.1)$$

where i = the random group,

n = the number of random groups,

$\hat{y}_i$ = the estimate of the category based on a specified random group,

$\hat{\bar{y}}$ = the mean of the specified category estimate.

The 2002 SBO used 10 random groups for the noncertainty cases and one random group for all certainty cases. Records were sorted in the same order as they were during sampling. Then, random groups were assigned consecutively from 1 to 10

for all noncertainty cases. Those cases selected with certainty were assigned automatically to random group 0 (Schlein, 2005).

The variance was modified to account for the imputed values of gender, ethnicity, and race for one or more owners of a company (Schlein, 2005). This variance adjustment factor was applied to the above variance equation. The variances of the predictor variables later used in the model fitting were then calculated using the same methodology as that of the original SBO data.

Chapter 4: Fixed-Effect Linear Model

4.1 Model Selection Techniques

The SAS programming language was used to help build a regression model that would fit the 2002 SBO data. In order to construct a model, the data could be fitted either at the unit level or at some aggregate level. Both methods were attempted in this study. At the unit (company) level, there were a great many predictor variables in the model and the R^2 values were very low. Hence, the data were aggregated to county level and a model was fit on that data. Significantly fewer predictors were included in the regression model at the county level.

Finding a model with as few variables as needed to maintain good predictability is important to finding a suitable Fay-Herriot model. Therefore, a linear regression model was fitted in order to screen for variables to use in the Fay-Herriot model. Aggregating the data to county level was determined to be the best way to fit a regression model, in this instance, due to the interest in small area estimation.

Initially, however, 2002 SBO company-level survey data were combined with administrative data by a unique company identifier. This data included race, gender, and Hispanic or Latino origin information provided in a previous SBO survey, in the decennial census, or in some other survey. (For a list of variable descriptions see Variable Glossary.) Then, responses for each variable were transformed into binary variables. Terms for inclusion into the model might

involve the main variables and also the squares, cross products, or other combinations and transformations of the principal variables (Draper & Smith, 1981). Hence, up to four way interactions of these variables were created and used in the modeling. Finally, three separate model selection techniques were used to find the best model: forward selection, backward selection, and stepwise selection.

SAS uses a default significance level of 0.5 for the forward selection, 0.15 for stepwise selection, and 0.1 for backward selection. To place greater restrictions on the number of variables allowed into the regression models, however, a level of 0.01 was chosen for all selection methods.

The forward selection technique begins with no variables in the model. For each independent variable, an F-statistic is calculated. The p-values for the F-statistics are then compared to a predetermined significance level, 0.01 for this study, adding the variable with the largest F-statistic significance level to the model (Draper & Smith, 1981). The process is repeated until no variable produces an F-statistic with significance level greater than 0.01 (SAS, 2003).

In the backward elimination procedure, a regression equation containing all variables is computed. Then, a partial F statistic is calculated for every predictor variable as if it were the last variable to be entered into the regression equation. The smallest partial F-test is then compared to a preset significance level, in this case 0.01. The corresponding predictor variable is removed if its partial F-test is

not significant at the 0.01 level and the process is repeated. Otherwise, the regression equation is adopted as calculated (Draper & Smith, 1981).

In the stepwise regression procedure, predictor variables are inserted into the regression equation until a satisfactory one is developed. The order of insertion is determined by using the partial correlation coefficient as a measure of the importance of variables not yet in the equation (Draper & Smith). Like the forward selection technique, variables are added one at a time to the regression equation provided their F-statistic is greater than the given significance level, again 0.01. After including a variable, the stepwise method looks at all the variables already included in the model and deletes any that do not produce an F-statistic greater than the chosen significance level. The process ends when none of the variables outside the model has a significant F-statistic and every variable in the model is significant at the SLENTRY level, 0.01, or when the variable to be added is the same one that was just deleted (SAS, 2003).

Using the entire 2002 SBO universe of 2.3 million businesses in the three types of model selection techniques would have required more computing resources than were available. So, in order to make computing more manageable, one state at a time was run through the regression model building process. Initially, it was desirable to have only one model that could be used for all states, if not across the nation, then at least in a few large regional groups. However, upon examination of the models for Georgia and Ohio, it was shown that vastly different variables were included in each of the models. Hence, some further study into the number

of models needed to cover the entire United States may be conducted in future work.

4.2 Data Exploration

The survey data were augmented with two types of auxiliary data. The first was previous survey data. The 1997 SMOBE/SWOBE race, gender, and Hispanic or Latino origin of companies that were selected again in the 2002 SBO were affixed to the dataset. However, only 0.3% of the 2002 sample was also included in the 1997 survey. Examination of this data showed that the previous survey data were not sufficiently numerous to have an impact on prediction.

The second type of auxiliary data was received from various administrative sources giving race, gender, and Hispanic or Latino origin. Nearly 20% of the businesses in sample contained such data. Inspection of the data showed that administrative records were present only for the nonemployer companies, thereby giving the impression that better prediction was possible for those businesses. The nonemployer companies had this administrative data because owners were able to be matched by SSN to other sources like the decennial census.

Due to the existence of these administrative records, the data were split into three subsets: one containing employer records, one containing nonemployer records with administrative data, and one containing nonemployer records without administrative data. The idea that the groups behaved so differently suggested that different models would be necessary.

Table 4.1 Number of Records in Subgroup Datasets

	Georgia		Ohio	
	Number	Percent of Total	Number	Percent of Total
Nonemployers with Administrative Data	18,384	22.7%	21,155	21.6%
Nonemployers without Administrative Data	8,299	10.3%	12,030	12.3%
Employers	54,260	67.0%	64,676	66.1%
Total	80,943	100.0%	97,861	100.0%

In the state of Georgia, when all records were grouped together, the results of the forward, backward, and stepwise model selection techniques all yielded models with 60 or more variables and R^2 values of 0.39 or less. On the contrary, the subset of that data containing only nonemployer records produced models with R^2 values of 0.76, while the subset containing only employer records produced models with R^2 values of 0.13. Each of these models, however, contained about 30 variables.

4.3 Special Data Handling

Due to the poor performance of the modeling at the unit level and the consideration of small area estimation techniques, this is the point at which the data were aggregated to the county level. This procedure involved recoding the characteristics of each business in the universe and the survey responses for each business in the sample to binary variables. Those indicators were then aggregated

by county. The county-level sums were then used with the estimated total number of businesses in each county in order to create a proportion of businesses by county with the given characteristic. These proportions, in turn, were transformed using the logit function.

Unfortunately, much of the data had missing values for several variables. Records in the universe that were not selected for sample had missing values for information obtained in the survey. Employers had missing values for owner-level information from administrative data. These variables were recoded with missing values changed to zero so that no variables were excluded from the analysis. Additionally, all character variables were converted to numeric values.

To determine whether the demographic make-up of a state or county had any bearing on the probability of a business being Black-owned, state and county population percentages were obtained. These proportions were then converted using the logit function. Because some of the county population proportion estimates were zero for certain races, the logit function was modified for all variables as follows:

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{C+0.5}{T-C+0.5}\right) \qquad (4.1)$$

where C is the characteristic of interest within a state (or county)

T is total population of a state (or county)

p is $\frac{C}{T}$, the proportion of the state (or county) with the given characteristic.

This adjustment ensured that no county-level variables were missing, thereby causing records to be deleted from the regression analysis.

4.4 *Initial County Models*

In performing the regression analysis, assumptions were made that the errors were independent, had zero mean, a constant variance, and followed a normal distribution. The residuals should exhibit tendencies that tend to confirm those assumptions, or at least should not refute the assumptions (Draper & Smith, 1981). An examination of the residuals showed that they did not necessarily follow a normal distribution. When viewing a histogram of the residuals, it was shown that the tails were not indicative of a normal distribution. Figure 4.1 shows the residuals for Georgia and Figure 4.3 displays the residuals for the state of Ohio. A quantile-quantile (q-q) plot (Figure 4.2) shows that the tails tend away from the normal distribution.

Figure 4.1 County-level Histogram of Residuals for Georgia

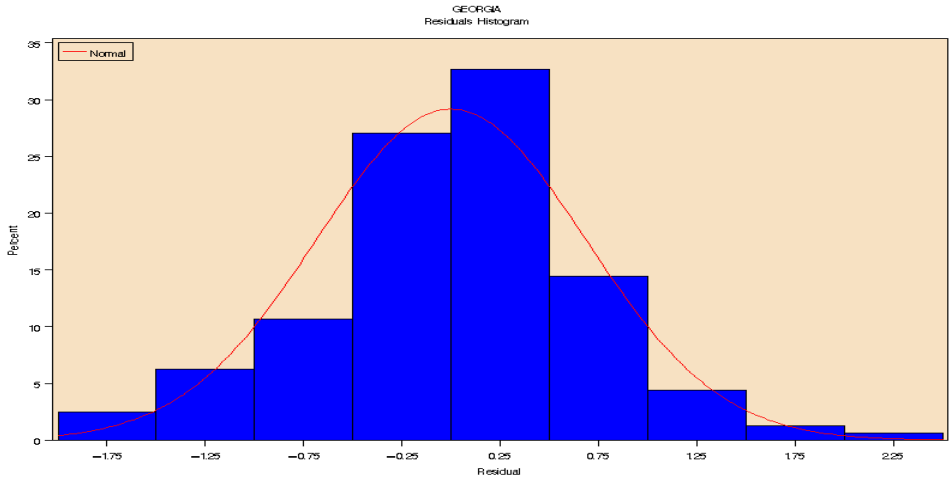


Figure 4.2 Q-Q Plot of Residuals for Georgia

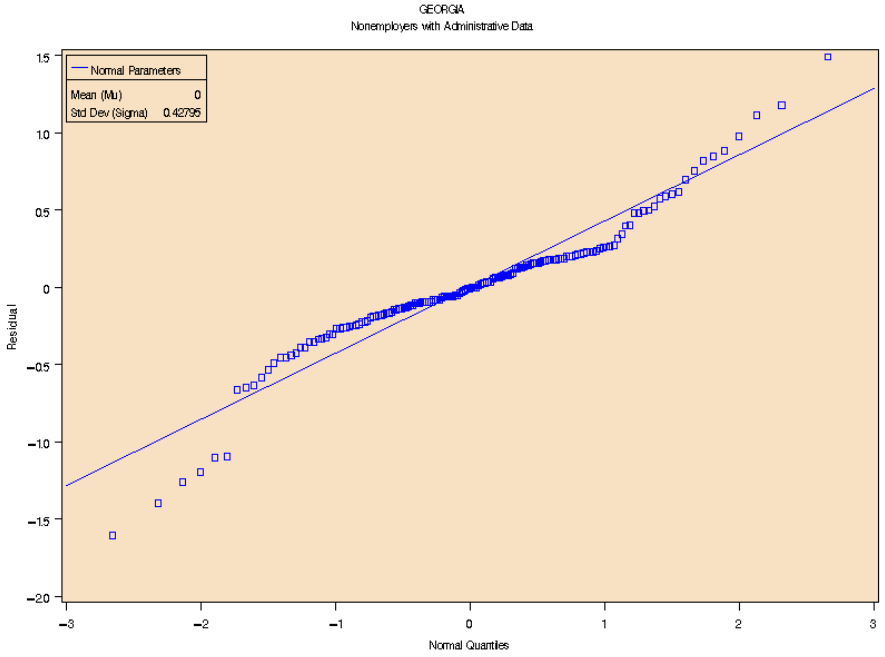
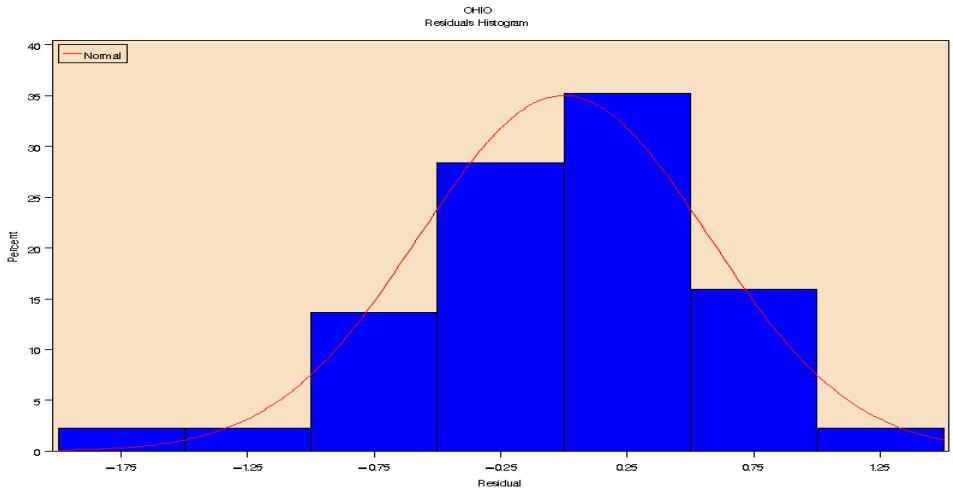


Figure 4.3 County-level Histogram of Residuals for Ohio



In order to obtain a suitable model with good predictability and few variables, it was apparent that some method other than the automatic variable selection techniques from Section 4.1 would be necessary. Automatic selection methods

were producing models with many variables and low R^2 values. To determine which variables had the greatest impact on predicting Black-ownership, the partial correlations of all available variables were examined. Though this method was similar to forward selection, it allowed for greater control over which variables were entered into the model. As stated in Draper & Smith (1981), the partial correlations can be found for the portions of the original data vectors, and they have no dependence on the values of the predictor.

SAS's PROC CORR was used to compute the Pearson correlation coefficient and probabilities for each analysis variable. They were added to the regression model if the predictor was the most correlated to Black-ownership after the effect of any previously added variables had been removed (Draper & Smith, 1981). In essence, this procedure allowed for obtaining partial correlations by measuring the strength of relationship between two variables while controlling for the effect of one or more others. When one variable was highly correlated with the probability of being Black-owned, it was added to the list of controlled variables. This process was continued until no other predictors had a 10% or greater partial correlation coefficient with the response variable. At that point, the list of variables was put into a regression model statement to obtain a value for R^2.

The variable found to have the greatest Pearson correlation coefficient with BLACK, the variable indicating a Black-owned business, was NCOBLACK, the proportion of a county's demographic population that is Black. When NCOBLACK was used as the first variable in the model, the next most highly correlated variable differed depending on which state was being examined. In the

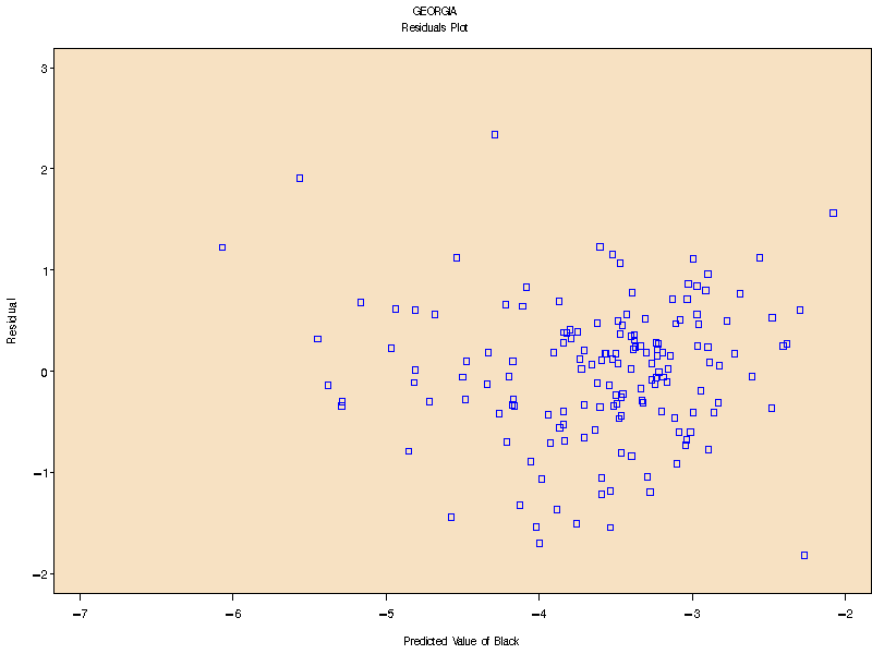
state of Georgia, the model obtained using this method had 13 variables and an R^2 value of 0.61. In Ohio, a model was produced containing only 8 variables with an R^2 value of 0.41.

The thirteen variables produced for Georgia were NCOBLACK, NCOLFONR, NCOWHITE, NCOSEC31, NCOASIAN, NCOLFONR*NCOWHITE, NCOSEC31*NCOASIAN, NCOTWO, NCOEMPSZ6, NCOSEC23, NCOEMPSZ3, NCOPOBFIN, and NCOPOBFIN*NCOEMPSZ3. (See Table 4.2 for variable descriptions.) Upon inspection of the model, several of the variables appeared to be somewhat insignificant. Many of these predictors had very low t-values and were therefore dropped from consideration. This produced a model with 4 variables and an R^2 of 0.53.

Table 4.2 Predictor Variables for Georgia

Predictors in Model	Description	Sign
• NCOBLACK	• County proportion of Black population	-
• NCOLFONR	• County proportion of businesses with LFO (legal form of organization) type not reported	-
• NCOWHITE	• County proportion of White population	-
• NCOSEC31	• County proportion of businesses in Sector 31-Manufacturing	-
• NCOASIAN	• County proportion of Asian population	-
• NCOLFONR*NCOWHITE	• Interaction of county proportion of businesses with LFO type not reported and county proportion White population	+
• NCOSEC31*NCOASIAN	• Interaction of county proportion of businesses in Sector 31 and county proportion Asian population	-
• NCOTWO	• County proportion of two or more races population	+
• NCOEMPSZ6	• County proportion of businesses with 2500+ employees	+
• NCOSEC23	• County proportion of businesses in Sector 23-Construction	+
• NCOEMPSZ3	• County proportion of businesses with 10-99 employees	+
• NCOPOBFIN	• County proportion of businesses with place of birth foreign indicator = 1	+
• NCOPOBFIN*NCOEMPSZ3	• Interaction of county proportion of businesses with place of birth foreign indicator = 1 and county proportion of businesses with 10-99 employees	+

Figure 4.4 County-level Residuals Plot for Georgia



The existence of outliers was evaluated through the inspection of plots of the residuals against the predictor variables. An examination of the residual plots did not indicate that the chosen predictors violated any assumptions made about the model. (See Figures 4.4 and 4.7.) In addition to this evaluation, also inspected was whether other potential predictors were needed in the model. The residual plots of other covariates against the residuals did not show that any of the other terms would be necessary in the model. Figures 4.5 and 4.6 show examples of residuals against covariates included and not included in the model, respectively.

Figure 4.5 County-level Plot of Residuals Against an Included Covariate (NCOEMPSZ3) for Georgia

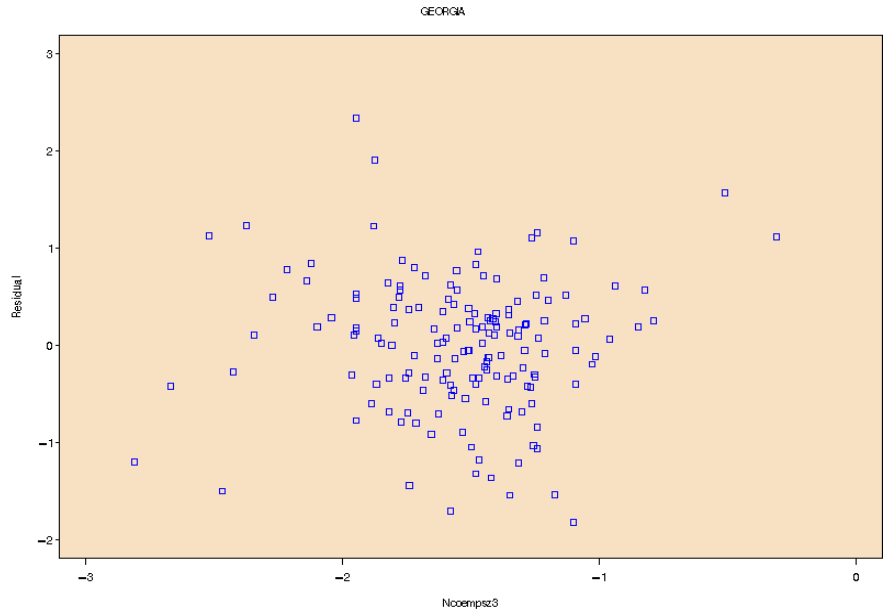
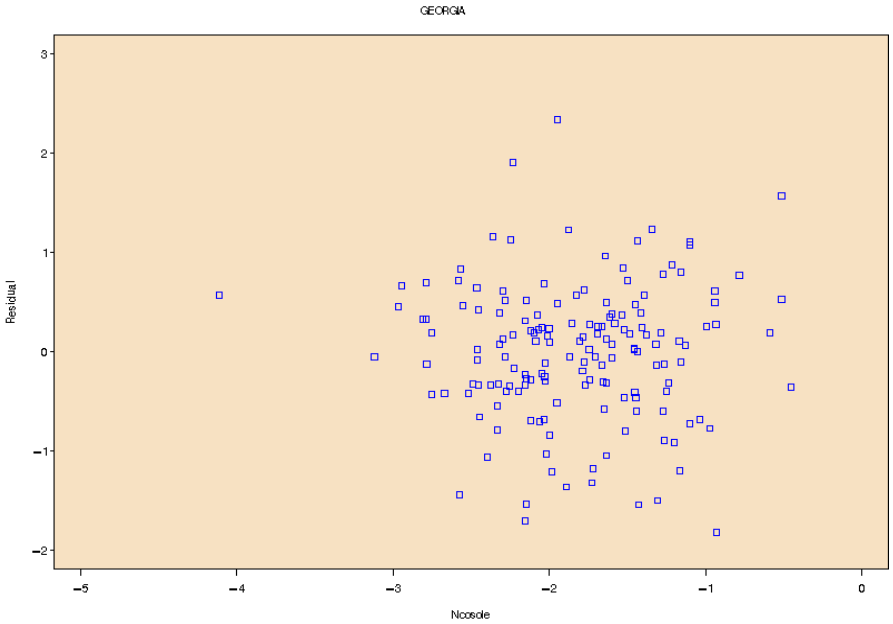


Figure 4.6 County-level Plot of Residuals Against an Excluded Covariate (NCOSOLE) for Georgia

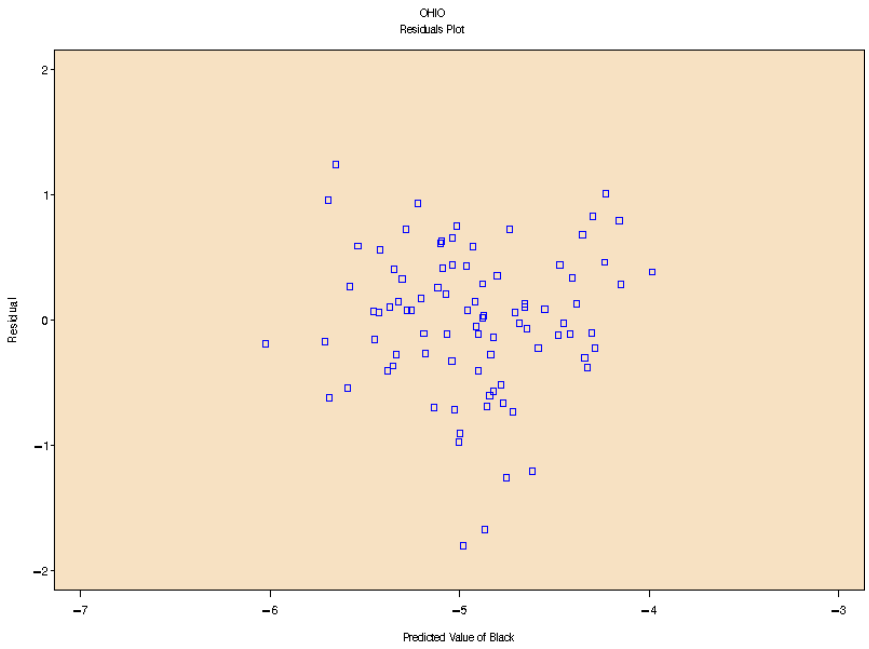


Eight variables that were used in the model statement for Ohio were NCOBLACK, NCOPOBFIN*NCOWHITE, NCOSEC21, NCOTWO, NCOSEC42, NCOSEC11, NCOPOBFIN, and NCOWHITE. (See Table 4.3 for variable descriptions.) Disregarding those that were not greatly significant produced a model with 3 variables and an R^2 of 0.39.

Table 4.3 Predictor Variables for Ohio

Predictors in Model	Description	Sign
• NCOBLACK	• County proportion of Black population	+
• NCOPOBFIN*NCOWHITE	• Interaction of county proportion of businesses with place of birth foreign indicator = 1 and county proportion of White population	+
• NCOSEC21	• County proportion of businesses in Sector 21-Mining	+
• NCOTWO	• County proportion of two or more races population	+
• NCOSEC42	• County proportion of businesses in Sector 42-Wholesale trade	+
• NCOSEC11	• County proportion of businesses in Sector 11-Agriculture, Forestry, Fishing & Hunting	-
• NCOPOBFIN	• County proportion of businesses with place of birth foreign indicator = 1	-
• NCOWHITE	• County proportion of White population	+

Figure 4.7 County-level Residuals Plot for Ohio



4.5 *Cross Validation of Initial Models*

To test the models that were developed using PROC CORR, the data were put through a cross validation trial. To examine the effectiveness of the models that were developed for Georgia and Ohio, a random sample of the data was selected. Approximately one-fourth of the data were chosen to be in the test set, with the remainder in the training set. Then, a regression model was run on the data, producing coefficients solely for the observations that made up the training dataset. For those training set observations, the sum of squares was calculated. This process was repeated 100 times using 100 different random samples. The mean sum of squares in Georgia was 19.058 and in Ohio it was 8.115.

When the regression was run on the entire dataset for Georgia, the total sum of squares was 151.07651. Using the following formula, a factor of reduction with an effective R^2 was calculated:

$$\frac{SSQ_{avg}}{N^*} = \left(1 - R^{*2}\right)\frac{SSQ_T}{N-1} \qquad (4.2)$$

where SSQ_{avg} = the average sum of squares from 100 samples,
SSQ_T = the total sum of squares,
N^* = the number of observations in the test dataset,
N = the total number of observations, and
$1\text{-}R^{*2}$ = the factor by which the variance can be reduced.

In Georgia, N was 159, N^* was 40, the R^2 was 0.5109, and the adjusted R^2 was 0.4982. Using the above formula, R^{*2} was found to be 0.5017.

In Ohio, the total sum of squares was 43.84301, N was 88, N^* was 22, the R^2 was 0.3558, and the adjusted R^2 was 0.3328. Using the above formula, R^{*2} was found to be 0.2681 for Ohio.

For each of these states, it was shown that the regression model chosen was a good model. Comparing the adjusted R^2 with the factor of reduction showed that there is not much difference between the two. Additionally, a histogram of the residuals shows that they resemble a normal distribution with mean zero. The residuals give the differences between what is actually observed and what is predicted by the regression equation (Draper & Smith, 1981).

4.6 Final Fixed-Effect Models

As mentioned in Section 4.2, a great advantage was to be gained by partitioning the county-level data into three subsets. The groups created were nonemployers containing administrative data, nonemployers without administrative data, and employers. Again, all variable responses were transformed into binary variables that were then aggregated to the county level. Using the model selection techniques from Section 4.1, a model was obtained. In most cases, the predictor variable set was pruned to have as few variables as possible without losing too much predictive power.

In Georgia, the nonemployers containing administrative data yielded an R^2 value of 0.86 having only one predictor, NCOADBLACK—the indicator of Black-ownership from an administrative record. Likewise, in Ohio, an R^2 value of 0.76 was produced from only NCOADBLACK. The presence of an administrative record appeared to be greatly predictive of being Black-owned. Clearly, the county-level model had much greater predictive power and many fewer variables than the unit-level model.

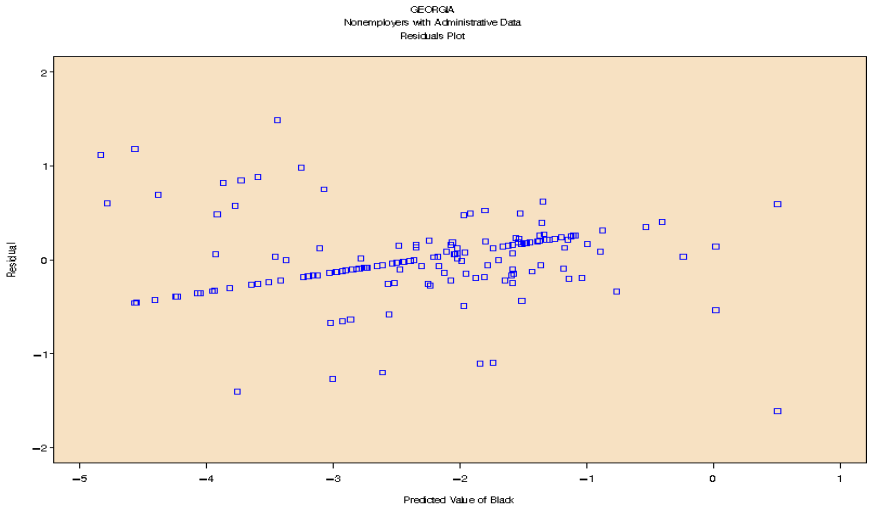
Table 4.4 Predictor Variables for Nonemployers with Administrative Data

State	Predictors in Model	Description	Sign
Georgia	• NCOADBLACK	• County proportion of indicators of presence of administrative data indicating Black-ownership	+
Ohio	• NCOADBLACK	• County proportion of indicators of presence of administrative data indicating Black-ownership	+

The formation of this variable, NCOADBLACK, was as follows. Three administrative sources each identified a race for the company's owner(s)—NARRACE, NBESTRACE, NCENRACE. They were then ordered by reliability and the first nonmissing value was used to create a variable for the administrative race—NRACE. Then, those values were transformed into indicators, which were aggregated to county level, and the modified logit proportion was formed.

In Georgia, the residuals plot (Figure 4.8) of the nonemployer businesses with administrative records did not appear to be random. Instead, it showed a linear relationship for counties where all sampled businesses had their Black-ownership correctly given by the administrative record. The weighted, estimated county-level NCOADBLACK predictor and the county-level BLACK response variable also agreed. The model with only NCOADBLACK as a predictor, therefore, appeared to contain a pattern of a perfect line embedded in the plot. This can be explained by the exact equality between Black-ownership indicators and administrative record race indicators for all businesses within a sizeable subset of Georgia's counties.

Figure 4.8 County-level Residuals Plot for Nonemployer Businesses with Administrative Data for Georgia



The data set of nonemployer businesses without administrative records in Georgia gave an R^2 value of 0.65 with five variables: NCOSEC21, NCOWHITE, NCOSEC21*NCOWHITE, NCOSEC31, and NCOSEC23. Similarly, in Ohio, the dataset of nonemployers without administrative data produced a model with an R^2 of 0.55 and five predictors: NCORETAIL, NCOSEC11, NCOWHITE, NCOPOBFIN, NCOSEC11*NCOWHITE.

Table 4.5 Predictor Variables for Nonemployers without Administrative Data

State	Predictors in Model	Description	Sign
Georgia	• NCOSEC21	• County proportion of businesses in Sector 21-Mining	+
	• NCOWHITE	• County proportion of White population	+
	• NCOSEC21*NCOWHITE	• Interaction of county proportion of businesses in Sector 21 and county proportion of White population	+
	• NCOSEC31	• County proportion of businesses in Sector 31-Manufacturing	+
	• NCOSEC23	• County proportion of businesses in Sector 23-Construction	+
Ohio	• NCORETAIL	• County proportion of retail business	+
	• NCOSEC11	• County proportion of businesses in Sector 11-Agriculture, Forestry, Fishing & Hunting	+
	• NCOWHITE	• County proportion of White population	+
	• NCOPOBFIN	• County proportion of businesses with place of birth foreign indicator = 1	+
	• NCOSEC11*NCOWHITE	• Interaction of County proportion of businesses in Sector 11 and county proportion of White population	+

The employer subgroup in Georgia produced a six-predictor variable model with an R^2 of 0.52. The predictors were NCOLFOOTH*NCOWHITE, NCOLFOOTH, NCOWHITE, NCOSEC31, NCOFAMY, NCOSEC31*NCOFAMY. In Ohio, the employer subgroup model had an R^2 value of 0.48 with five predictor variables: NCOSEC22, NCOHISP, NCOEMP, NCOWHITE, and NCOEMP*NCOWHITE.

Table 4.6 Predictor Variables for Employers

State	Predictors in Model	Description	Sign
Georgia	• NCOLFOOTH*NCOWHITE	• Interaction of county proportion of businesses with LFO type other and county proportion of White population	+
	• NCOLFOOTH	• County proportion of businesses with LFO type other	+
	• NCOWHITE	• County proportion White	-
	• NCOSEC31	• County proportion of businesses in Sector 31-Manufacturing	+
	• NCOFAMY	• County proportion of family-owned business	+
	• NCOSEC31*NCOFAMY	• Interaction of county proportion of businesses in sector 31and county proportion of family-owned business	+
Ohio	• NCOSEC22	• County proportion of businesses in sector 22-Utilities	-
	• NCOHISP	• County proportion Hispanic	-
	• NCOEMP	• County proportion of employers	+
	• NCOWHITE	• County proportion White	+
	• NCOEMP*NCOWHITE	• Interaction of county proportion of employers and county proportion White	-

Once the coefficients for the models were obtained, the variances were calculated for the predictor variables under the same methodology as published SBO data, described in Section 3.6. This was done because the sampling variances are assumed known when using the Fay-Herriot estimation method.

Chapter 5: Small Area Results

5.1 Fitting the Fay-Herriot Model

The Fay-Herriot model given in Equation (2.1) was fit using the predictors determined in the linear model regression of Section 4.6. Again, there were three models per state—one each for nonemployers with administrative data, nonemployers without administrative data, and employers.

The dependent variable consisted of the weighted, modified logit-transformed Black proportion of businesses by county, FH_BLACK. This variable was estimated using Equation (4.1) with C = estimated number of Black-owned businesses in the county and T = estimated number of all businesses in the county. The variance for this variable was estimated using a delta-method approximation given by

$$\hat{\sigma}^2_{FH\_BLACK} = \frac{\hat{\sigma}^2_C}{(C+0.5)^2} + \frac{\hat{\sigma}^2_T}{(T+1)^2} - \frac{\hat{\sigma}^2_{C+T} - \hat{\sigma}^2_C - \hat{\sigma}^2_T}{(C+0.5)(T+1)} \qquad (5.1)$$

where C = estimated number of Black-owned businesses per county,

T = estimated number of all businesses per county,

$\hat{\sigma}^2_C$ = estimated variance of C,

$\hat{\sigma}^2_T$ = estimated variance of T, and

$\hat{\sigma}^2_{C+T}$ = estimated variance of estimated (C+T) variable.

The variances above were derived from the data using the random groups technique described in Equation (3.1). A delta method approximation was chosen as a suitable way to compute the variance for the estimated proportion of Black-owned businesses.

Although the sampling error variances are often assumed known, it is common practice—especially within the Census Bureau—to estimate such variances using generalized variance functions (GVFs). These functions are used to approximate the design-based variance estimators of target means and proportions. A GVF is a mathematical model that describes the relationship between the variance of a survey estimator and its expectation (Wolter, 1985). As compared to design-based variance estimators computed directly from survey microdata, GVFs have several advantages including operational simplicity, increased stability of standard errors, and reduction of disclosure limitation problems for cases involving public-use datasets (Eltinge, Jang, & Choi, 2002).

The estimated sampling error variance, by county, for each subgroup was taken to be

$$\hat{s}_{i,e}^2 = \frac{V_e}{n_{i,e}} \qquad (5.2)$$

where V_e = constant (GVF) for subgroup e, and

$n_{i,e}$ = sample size of county I for subgroup e.

The constant, V_e was obtained by estimating $\log(V_e)$ as the average of $\log\left(\hat{\sigma}^2_{FH\_BLACK}\right)+\log(n_i)$ over all counties within each subgroup of data (i.e. nonemployers with administrative data, nonemployers without administrative data, and employers) and exponentiating that mean. Those constants are given in the table below.

Table 5.1 GVFs used in Calculating Sampling Error Variances

	Georgia	Ohio
Nonemployers with Administrative Data	13.00	44.50
Nonemployers without Administrative Data	3.75	5.40
Employers	15.00	35.00

Applying the predictor variables that were selected in the fixed-effect linear model and treating the sampling error variances as known, the Fay-Herriot model was fitted. Finally, the Fay-Herriot predictors were created using the EBLUP formula below.

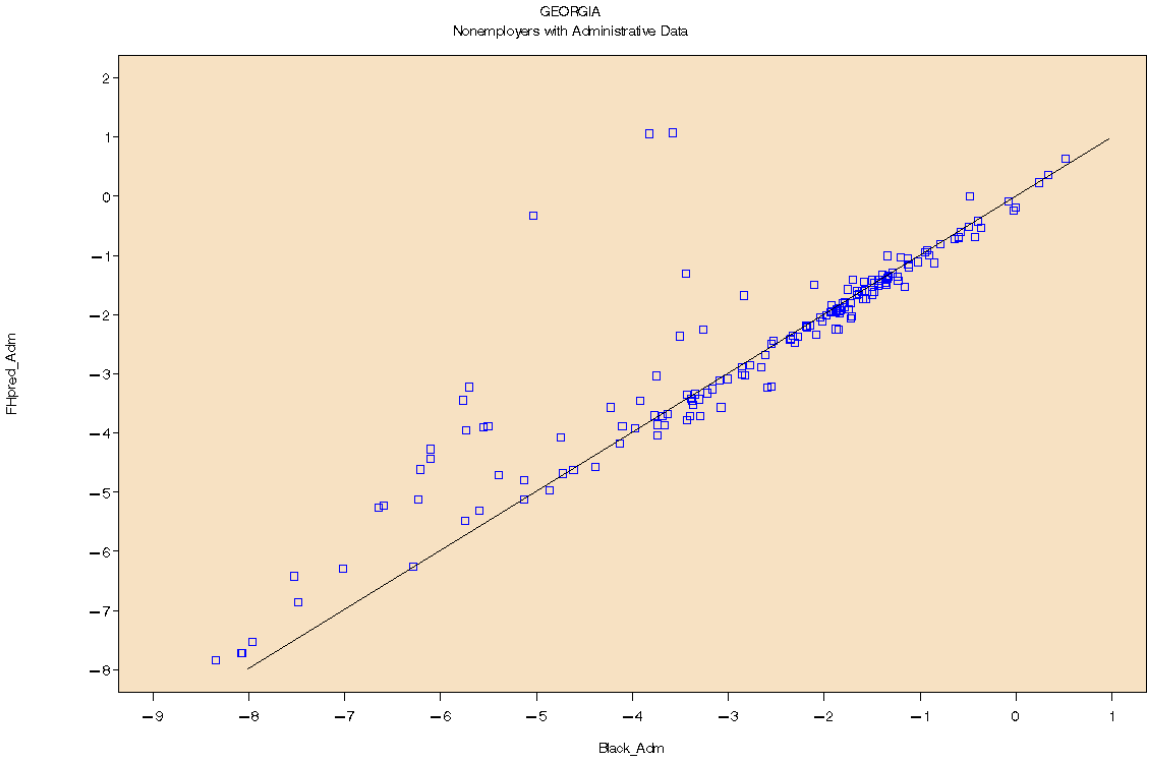
$$\tilde{\theta}_i = \frac{\sigma^2}{\sigma^2 + \frac{V_e}{n_{i,e}}} FH\_BLACK_i + \frac{\frac{V_e}{n_{i,e}}}{\sigma^2 + \frac{V_e}{n_{i,e}}} x_i^T \hat{\beta} \qquad (5.3)$$

5.2 Fay-Herriot Predictors

The EBLUP predictors produced by the Fay-Herriot small area estimation model showed that for very negative values of Black-ownership, the Fay-Herriot

predictions increase the proportions. This can be seen in Figure 5.1, which displays how the Fay-Herriot predictors correspond to the fixed-effect linear model predictors for the nonemployers with administrative data subset in Georgia. The larger values of the modified, logit-transformed Black-ownership percentage tend to decrease with the small area estimators.

Figure 5.1 Plot of Fay-Herriot Model Predictors Against Fixed-Effect Model Predictors for Nonemployers with Administrative Data in Georgia

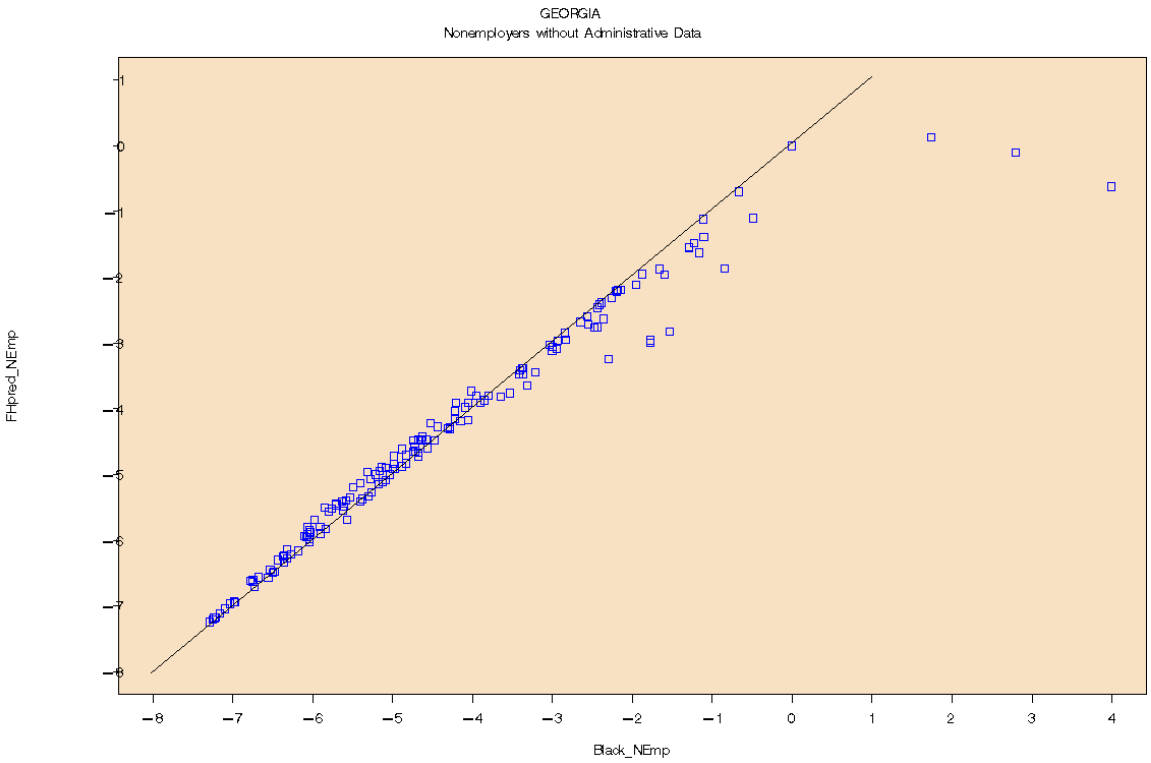


It is evident in Figure 5.1 that three counties in particular are adjusted drastically upwards from the estimates obtained through the fixed-effect modeling. These counties had particularly small sample sizes of 1 business each. However, as stated in Section 4.2, the fixed-effect linear model was fairly impressive in its

prediction for the subset of data containing nonemployers with administrative data. It is reasonable, then, to use the fixed-effect model for that subset and to use the Fay-Herriot prediction for the remaining two subgroups.

In Figure 5.2, for instance, there appears to be a particularly close association between the two predictors, with the exception of a few counties with larger proportions of Black-ownership. These three counties have small sample sizes of 1, 2, and 4 businesses. The fact that these counties fall among the tail end of the estimates may also suggest that some bias correction is needed. Recall that the estimation was done using modified, logit-transformed proportions. By taking the exponential of those proportions and bias correcting for the nonlogit-transformed predictors, these "outlier" counties may result in reasonable perturbations of the corresponding direct estimate to Fay-Herriot predicted proportions of Black ownership.

Figure 5.2 Plot of Fay-Herriot Model Predictors Against Fixed-Effect Model Predictors for Nonemployers with Administrative Data in Ohio



At present, there is no complete and accurate source of Black-owned business data to cross validate the results of the small area prediction. However, this study showed that the borrowed strength of the Fay-Herriot estimates produces very sensible results.

Chapter 6: Conclusions

6.1 Summary

As expected, use of the Fay-Herriot model predictors gave seemingly reasonable estimates of county-level proportions of Black business ownership. The effects of "smoothing" the data show that the predictions do not differ tremendously from the fixed-effect linear model predictions. Only in certain cases, often in counties with extremely small sample sizes, did the Fay-Herriot estimates vary greatly from the raw estimates.

The use of auxiliary data proved to be extremely important in the predictions. The mere presence of auxiliary data, in particular the administrative record indicating Black-ownership, was so predictive that it warranted disaggregating the data into three separate subgroups. In fact, the one variable gave such strong predictability that it suggests the raw estimates may be sufficient for that subset of data.

The small area methodology does very well for the nonemployers without administrative data and the employers subgroups. It would make sense, therefore to reaggregate these subsets to get overall predicted county-level proportions of Black ownership.

All variables selected in the various subgroups' models are available for non-sampled counties as well. This was an unexpected benefit found in this study. As

a result of the availability of all predictor variables, the small area estimation should have much greater predictive power in those small areas that are not sampled. Those areas not sampled will be able to borrow strength from other areas using the same predictors.

6.2 *Future Work*

The scope of this research covered only one of the many characteristics that the Survey of Business Owners and Self-Employed Persons is interested in estimating, Black-ownership. Future work should encompass all of the main traits observed in the survey, including all race, gender, and ethnicity ownership characteristics. Additionally, when the modified, logit-transformed predictors are obtained, they would need to be either exponentiated or transformed by the logistic distribution function. Therefore, some appropriate bias correction should be done on the estimates. It is highly likely that doing so would produce very reliable estimates of the proportions of ownership by characteristic. The estimated *MSE*s for the small-area estimates could also be compared to those for the direct estimates in order to gauge their reliability.

A small study of Hispanic-ownership was done during this research study. The linear fixed-effect prediction before any subsetting of the data appeared to have very low prediction capabilities. In Georgia, the unit-level model fitting produced a model with 6 variables and an R^2 value of 0.48. At the county level, however, the Hispanic-owned model contained 5 variables and had an R^2 value of 0.38.

Those variables were NCOHISP, NCOFRAME, NCOPOBFIN, NCOSEC42, and NCOTWO.

The model for Ohio did not appear to be very predictive of a business being Hispanic-owned. The unit-level model contained 5 variables and had an R^2 value of 0.36. However, once aggregated to the county-level, the model had 5 variables predictive of being Hispanic-owned. The R^2 given was 0.16. Contrary to the predictive power of the models for Black-ownership, the Hispanic-ownership models at the unit level were better than the county-level aggregate models.

It seems, then, obvious that some investigation into whether partitioning the data in the three subgroups would have as great an impact on the prediction of Hispanic-owned businesses as it did on Black-owned businesses. Such research would prove to be quite useful to the SBO staff, as estimates for this sub-domain tend to be underestimated.

Overall, the use of Fay-Herriot estimates appears to be reasonably effective. Some future work encompassing the identification of the number of different models necessary, the level of modeling that should be done, and the availability of predictors for non-sampled small areas is of great interest. Such research could lead to a remarkable improvement of estimates in the SBO.

Appendices

Variable Glossary

Variable Name	Description
NBlack	Logit of the proportion of businesses indicating Black ownership (response variable)
Ncoadblack	Logit of the proportion of businesses with administrative record indicating Black ownership
Ncoadother	Logit of the proportion of businesses with administrative record indicating some other race ownership
Ncoadwhite	Logit of the proportion of businesses with administrative record indicating White ownership
Ncoblack (or Ncoasian, etc.)	Logit of the proportion of county population that is Black (or Asian, etc.)
Ncocorp	Logit of the proportion of businesses with LFO (legal form of organization) type corporation
Ncoempsz1	Logit of the proportion of businesses with no employees
Ncoempsz2	Logit of the proportion of businesses with 1 to 9 employees
Ncoempsz3	Logit of the proportion of businesses with 10 to 99 employees
Ncoempsz4	Logit of the proportion of businesses with 100 to 999 employees
Ncoempsz5	Logit of the proportion of businesses with 1000 to 2499 employees
Ncoempsz6	Logit of the proportion of businesses with 2500 + employees
Ncofam1	Logit of the proportion of the one-owner businesses in the county
Ncofamy	Logit of the proportion of the family-owned businesses in the county
Ncoframe	Logit of the proportion of businesses in the Black frame
Ncofranchy	Logit of the proportion of the franchised businesses in the county
Ncohisp	Logit of the proportion of businesses tabbed as Hispanic
Ncohispanic	Logit of the proportion of businesses with administrative record indicating Hispanic ownership
Ncohome	Logit of the proportion of the homebased businesses in the county
Ncolfonr	Logit of the proportion of businesses with LFO type not reported
Ncolfooth	Logit of the proportion of businesses with LFO type other
Ncopartner	Logit of the proportion of businesses with LFO type partnership
Ncopobfin	Place of birth foreign indicator
Ncoretail	Logit of the proportion of businesses in retail
Ncosec11 - Ncosec99	Logit of the proportion of businesses in sector 11 - 99 (see Appendix B for all sectors)

Variable Name	Description
Ncosole	Logit of the proportion of businesses with LFO type sole proprietorship
Ncounty	County code
Ncowhite	Logit of the proportion of businesses tabbed as White
Ncowoman	Logit of the proportion of businesses tabbed as female
Nstasian (or NstBlack, etc.)	Logit of the proportion of state that is Asian (or Black, etc.)
Sbowgt	Weight
Tab_sboid	Unique SBO identifier

NAICS Sector Codes

Sector	Name
11	Agriculture, Forestry, Fishing and Hunting
21	Mining
22	Utilities
23	Construction
31	Manufacturing
42	Wholesale Trade
44	Retail Trade
48	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration
99	Unclassified

Bibliography

Census Bureau. (3/22/2007). [WWW page]. URL www.census.gov/csd/sbo.

Draper, N.R. & Smith, H. (1981). Applied Regression Analysis (2nd edition). New York, NY: John Wiley & Sons.

Executive Office of the President. (2002). North American Industry Classification.

Eltinge, J.L., Jang, D.S., & Chol, M.J. (2002). Use of Generalized Variance Function Models in Inference From Social and Economic Survey Data. Proceedings of Statistics Canada Symposium 2002.

Fay, R. and Herriot, R. (1979). Estimates of Income For Small Places: an Application of James-Stein Procedures to Census Data. Journal of American Statistical Association. 269-277.

Galvin, L. (2006). 2002 SBO—Assignment of Partnership and Corporation to Minority or Non-Minority Frames—Phase I.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, Vol. 9, No. 1 Feb 55-76.

Kacker, R.N. and Harville, D. A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models. Journal of American Statistical Association. 853-862

Longford, N. (2005). Missing Data and Small-Area Estimation. Leicester, England: Springer Science + Business Media, Inc.

Rao, J.N.K. (2003). Small Area Estimation. Hoboken, NJ: Wiley-Interscience.

SAS. (2003). SAS OnlineDoc 9.1. [WWW page]. URL http://support.sas.com/91doc

Schlein, B. (2005). [Internal Census Documentation]. 2002 SBO Estimation.

Williams, A. (2005). [Internal Census Documentation]. Item Imputation for SBO Nonresponse Cases.

Wolter, K. (1985). Introduction to Variance Estimation. New York, NY: Springer-Verlag.