Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Sarah E. Robertson Date

Comparison of Variance Estimations of Logistic Regression Models Produced by SAS, SUDAAN, and WesVarPC

By

Sarah Robertson

MPH

Emory University

Rollins School of Public Health

Department of Biostatistics

_____ [Chair Signature]

Paul S. Weiss

Thesis Advisor

_____ [Member Signature]

George A. Cotsonis

Comparison of Variance Estimations of Logistic Regression Models Produced by SAS, SUDAAN, and WesVarPC

By

Sarah E. Robertson

B.S.

Washington and Lee University

2010

MPH

Emory University

Rollins School of Public Health

2012

Thesis Committee Chair: Paul S. Weiss, M.S.

An abstract of

a thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Masters of Public Health

in Biostatistics

Abstract

Comparison of Variance Estimations of Logistic Regression Models Produced by SAS, SUDAAN, and WesVarPC

By Sarah E. Robertson

For this analysis, logistic regression models were built in SAS, SUDAAN, and WesVarPC using complex survey data with categorical variables. While the validity of models is important, this analysis focused more on the differences between the variance estimations produced by the three statistical software packages for logistic regression models. This analysis used a dataset collected by the Emory University Preparedness and Emergency Response Research Center, which provided data on Local Health Departments and their implementation of Incident Command Systems and Emergency Operation Centers. All the packages produced the same β estimates, except in the special case when there were zero observations in a level of an observed variable. Of the three packages, SUDAAN consistently produced the smallest standard errors, followed by SAS, and then WesVarPC.

Comparison of Variance Estimations of Logistic Regression Models Produced by SAS, SUDAAN, and WesVarPC

By

Sarah E. Robertson

B.S.

Washington and Lee University

2010

MPH

Emory University

Rollins School of Public Health

2012

Thesis Committee Chair: Paul S. Weiss, M.S.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

In partial fulfillment of the requirement for the degree of

Masters of Public Health

in Biostatistics

2012

Acknowledgements

I would like to thank the whole of the faculty, staff, and advisors of the Biostatistics Department at Rollins School of Public Health for facilitating my past two years of graduate study. I would also like to give a special thanks to four individuals at Rollins, who made this thesis possible. First, I would like to thank Elizabeth Bilter who had previously cleaned and weighted the dataset. I also wish to thank Michele Mindlin who allowed me to work with the data collected by the Emory University Preparedness and Emergency Response Research Center study, offered sage advice, and served as a reader for this paper. Thanks also go to Professor George Cotsonis for reading my thesis. I would like to express my gratitude to Professor Paul Weiss for his guidance throughout my career at Rollins, especially for acting as my advisor for this culminating project.

I also want to thank my friends and family; without their love, support, and encouragement, I would not have arrived at this point. Finally I would like to thank my mother for the many hours she spent as my final editor, correcting grammar and spelling on countless papers throughout my academic career.

Data used in this thesis was collected through a project supported by grant # 5P01TP000300 from the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC.

Table of Contents

Background

Introduction

The Institute of Medicine's 1988 report, *The Future of Public Health*, "defines the mission of public health as fulfilling society's interest in assuring conditions in which people can be healthy" (National Research Council). The incidents of 9/11, "anthrax (2001), SARS (2003), Hurricane Katrina (2005), and renewed concerns about pandemic influenza" (Koh, et al, 2008) helped to spur the Public Health field to incorporate emergency preparedness into its core mission and essential function. With federal funding allocated to the states to develop an emergency response infrastructure (Laurie, et al, 2006), many local health departments (LHD) initiated new emergency preparedness activities for the purpose of establishing response capacity and capabilities. These included the implementation of an Incident Command System (ICS) and/or an Emergency Operations Center (EOC).

Local Health Departments

The structure of LHDs differs widely from state to state. The National Association of City and County Health Officials (NACCHO), a membership group of LHDs within the United States, has categorized four different systems of organization: centralized, decentralized, shared, and mixed (FSRC, 2006). In the centralized organization, the LHD is under the sole control of the state; in the decentralized organization, the LHD is under the control of the local government, while the shared and mixed model systems allocate control to both state and local organizations (FSRC, 2006). Just as the structure of LHDs varies by state, the number of LHDs also varies by state.

Massachusetts, New Jersey, and Connecticut all have large numbers of LHDs relative to their geographic size and the population, while Rhode Island has none.

Incident Command System

In recognition on an increased emphasis on emergency preparedness since 9/11 in the United States, a federal mandate directed the adoption of the National Incident Management System (NIMS) in LHDs (Cox, 2011). ICS is the framework used by NIMS for emergency response, thus ICS became the framework for emergency response in the public health field. The establishment of ICS in public health has been challenging due to the differences between the traditionally collaborative field of public health and the urgency and hierarchical structure of emergency response organizations (Scott, 2010). ICS was established in the 1970s in the aftermath of devastating wildfires in California in 1970 (Irwin). Following this original disaster, the United States Forest Service analyzed problems encountered by the responding emergency services. They identified six major problems for the new system to address: interagency organization, communication, planning, information flow, management, and forecasting of future events (Irwin, 2000). In response to this analysis, the ICS was developed. An advantage to ICS is flexibility for each situation to which it is applied. An ICS is defined as, "a set of personnel, policies, procedures, facilities, and equipment, integrated into a common organizational structure designed to improve emergency response operations of all types and complexities" (Irwin, 2000). ICS is a hierarchical structure designed to respond to both large and small-scale emergencies of varying duration (Laurie, et al, 2006). This ICS framework, seen in Appendix E, consists of five specific components that can employ in total as many as 5,200 people. Incident Command serves an executive role in directing

the disaster response and is headed up by the Incident Commander. The Command Staff may also include a Safety Officer, Public Information Officer, as well as a Liaison Officer. The four Section Chiefs of Operations, Planning, Financial, and Logistics report to Incident Command and are charged with managing the tactical response of their sections (Hecht, et al, 2005). Each of these sections can consist of multiple levels of staff, dictated by the scale of the incident (Wikipedia). If the incident is small enough in scale, the Incident Commander can assume the duties of any other Section Chief.

A key advantage of ICS is the ability to smoothly coordinate emergency responses from different agencies following an incident. ICS is designed so that, for example, law enforcement, hospitals, public health, and other agencies are assigned clear roles and responsibilities under a single command structure. Few LHDs had a formal plan in place for emergency and incidence response before the federal mandate for NIMS. To help develop and implement ICS, many LHDs employed non-public health staff with previous experience in emergency response and ICS. This helped ease the introduction of a more "militaristic" and "hierarchical" ICS structure, that was used in emergency preparedness and response activities. This approach initially encountered pushback by some LHD leaders and public health staff (Laurie, et al, 2006).

Emergency Operations Center

Unlike the mandate requiring a LHD to implement an ICS, there is no federal mandate for a LHD to create an EOC. While ICS is a planned and non-physical command scheme, an EOC is an actual physical location used as the emergency response command center. Communication between EOC and the ground level responders is of utmost importance. Adequate information flow to and from the EOC may determine

whether and emergency response is a success or failure. The EOC is easily integrated into the ICS model as a physical command center where managerial decisions, logistics, and communication are based. EOCs can range from a minimal set-up of a table and computer(s) in one room that has been repurposed, to a complex, designated use space with an "integrated design" as seen in Appendix F (Cox, 2011). The EOC may be equipped with workstations, computers, phones, copiers, tables, TV monitors, radios, and similar equipment.

Survey

The data used for this analysis was taken from one phase of a large study using mixed methods, conducted by the Emory University Preparedness and Emergency Response Research Center (Emory PERRC) and spanning from May 2011 to February 2012. The study was conducted in three phases. Phase 1 consisted of phone interviews with "emergency preparedness coordinators" about ICS and EOC implementation. Phase 2 was a web-based survey of LHDs that focused on ICS, EOC, and emergency preparedness, with data collected from May 2011 to October 2011. The last phase, phase 3, consisted of focus groups (Mindlin, et al, 2011). The survey in phase 2 was initially constructed as a single-stage design with unequal probability of selection for primary sampling units (PSUs). After the initial sample was selected, issues arose with the sampling in the state of Georgia, which necessitated the use of the complex clustering in the final dataset.

SAS 9.3

For this analysis three statistical packages were used to produce variance estimations using the same complex survey data. Version 9.3 SAS is one of these

packages. SAS has not always been used to analyze complex survey data. The *proc* statements assumed a simple random sample until SAS Version 8.0 (Chen, 2004). The introduction of such procedures as *proc surveylogistic* in recent versions has made it possible to analyze complex survey data in SAS. SAS uses Taylor Series linearization for these specialized survey procedures.

SUDAAN 10

Survey Data Analysis, or SUDAAN, is a statistical software package designed specifically to address complex survey data that is not necessarily from a simple random sample design. There are two versions of SUDAAN, one which functions independently of SAS and one which is SAS-callable. For this analysis, the SAS-callable version 10 SUDAAN was used. Like the procedures used in SAS, SUDAAN uses Taylor series expansion linearization to create variance estimates and ratio estimators. The option to use replication methods of balanced repeated replication (BRR), jackknife, and bootstrap are also available, although these approaches are more easily used in other software packages.

WesVarPC 5.1

WesVarPC 5.1 is a statistical package that offers another method for complex survey data analysis. Unlike SAS and SUDAAN that utilize Taylor series expansion, WesVarPC relies solely on replication methods for variance estimation. The BRR method was used to create the variance estimates for this data. The jackknife and bootstrap methods are other options available with WesVarPC that use replication techniques.

Methods

Introduction

The survey was conducted by the Emory PERRC in the summer and early fall of 2011. This analysis used the web-based survey data from phase 2 of the study. By using a list of LHDs provided by NACCHO, a complex sampling method was used to ensure representation across the United States. The SAS, SUDAAN, and WesVarPC statistical packages were used to produce variance estimations. A predictive model for whether a LHD has implemented ICS was built into the packages, as well as a predictive model for whether a LHD has established an EOC in all three statistical packages.

Survey Design

The sampling frame of 2565 LHDs provided by NACHHO allowed a random sample of 704 to be drawn. Of these 704 LHDs, 696 were considered eligible sampling. The sampling fraction is defined as the equation below.

(1) $f = \frac{n}{N}$, where N = population size, n = sample size

For this original sample, the sampling fraction is equal to 0.2713, which is relatively small and does not indicate uncertainty. This finite population correction factor is defined in (2) as one minus the sampling fraction.

(2) $\text{Finite Population Correction} = 1 - f$

For the dataset, the finite population correction factor is equal to 0.7287. These values allow population and sample variances to be calculated. Equation (3) shows the sample variance calculation.

(3) $Var(y) = (1 - f)\frac{s^2}{n}, where\ s^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}\ and\ \overline{y} = \frac{\sum_{i=1}^{N} y_i}{N}$

The complex nature of the sample arises from the fact that the number of LHDs differed by state, requiring that some states be under-sampled, while others needed to be fully included. Four cities, Washington D.C., Los Angeles, New York, and Chicago, were selected with certainty, a probability of selection of 1, to ensure they were surveyed. This was done because each of these LHDs represents such a large portion of the population of the United States, as well as a unique LHD in terms of direct funding and the amount of money involved. If these LHDs were not in the sample, the survey would not be an overall representative sample of the United States. LHDs from Massachusetts, Connecticut, and New Jersey were under-sampled due to the high concentration of LHDs relative to the population represented and the total number of LHDs in the entire United States. Another issue with weighting arose for the state of Georgia since the original LHD list from NACCHO provided information by health districts, not by individual LHDs as indicated. This was discovered after the initial sample of LHDs was drawn. As a result, the initial random sample from Georgia was not of LHDs, but instead of health districts consisting of multiple LHDs. Therefore, LHDs were then randomly selected from the previously selected health districts, resulting in a two-stage survey design (Mindlin et al, 2011).

Weighting of Data

There were 441 observations and 335 variables in the original cleaned and weighted dataset. The weighting of the data is of extreme importance in this analysis. The final weight of the data is based on the sample design. A sample weight is defined as the inverse of the probability of selection.

(4) $w_i = \frac{1}{\pi_i}$ w_i = weight and π_i = probability of selection

The initial probability of selection is the inverse of the weight.

(5) $\pi_i = \frac{1}{wi}$ w_i = weight and π_i = probability of selection

For the survey, the LHDs of the cities of Los Angeles, Chicago, Washington D.C, and New York were giving an initial probability of selection equal to one. Due to non-response, the weights were readjusted. This is done by multiplying the sample weights by the inverse of the response rate (Heeringa).

(6) $W_i = w_i * rate_{resp} = w_i * \frac{\#\, surveyed}{\#\, responded}$

For this analysis five classes were used for this adjustment. The first class contained the cities selected with certainty. Of the LHDs selected with certainty, only two responded, which created an adjusted weight of 1.5 after the response rates were calculated. The class containing LHDs in New Jersey, Massachusetts, and Connecticut ended up with an adjusted weight of about 40.184. The final class consisted of all other states. The discovery that the LHD list provided by NACCHO was classified by health district in the state of Georgia, not by LHD, created the need for a second stage of weighting. This second adjustment was done only for the state of Georgia.

Primary Sampling Unit

Analysis through the three statistical packages of interest required that two forms of coding be used in terms of strata and PSUs. For the runs through SAS and SUDAAN, the stratification and clustering is that of the original survey design. This original design consists of fifty-two strata; one each of the two self-representing LHDs, and one for each state. Clustering is seen in the state of Georgia due to the initial sampling of health districts, not LHDs, which is accounted for in a clustering variable. Therefore the PSU variable for Georgia is the health district.

For the WesVarPC BRR variance estimations, a different way of coding for strata and PSUs was necessary. This was done by the creation of a total of 211 strata, in order to have two observations in each stratum. Two of the strata are self-representing. In order to create the other strata, the LHDs were arranged by the population each represents. Then a separate PSU was assigned to every pair of LHDs, totaling 209 strata. The LHDs in each stratum were each assigned the number 1 or 2 to represent the PSU.

Variable Selection

Multiple types of variables, including both categorical and nominal, were part of the original dataset. After examination of the 355 variables, twelve were considered for potential predictor variables for this analysis. The corresponding survey questions are:

1) *Number of individuals currently working for your local health department (include ALL full-time, part-time, and contractual employees).
2) *Which best describes the area served by your local health department?
3) *Does your local health department function as part of the state health agency?
4) Is the relationship with the state health agency for policy and program decisions different for emergency preparedness than for other health department activities?
5) To what extent does your state health agency specify emergency planning and preparedness requirements?
6) *Does your health department have a department or unit dedicated to emergency preparedness?

7) *How much funding for emergency preparedness did your health department receive in the last fiscal year? Include all sources of funding.

8) *Other than H1N1, how many events and incidents has your health department responded to since January 2009 (excluding drills and exercises)?

9) If an emergency occurred, would your health department have a formal role in the community response?

10) When did the local health department first respond to an H1N1 situation?

11) How does your local health department's leadership feel about ICS/EOC?

12) How does your local health department's non-emergency preparedness team feel about ISC/EOC?

* indicates selection for final model

After further examination in SAS, six of these variables, noted above, were selected as suitable variables for the analysis. These were classified as suitable variables due to their large response rates, relatively good distribution among responses, and potential predictive value. Two variables, the number of workers and amount of funding, were recoded to create fewer levels of response for each. This ensured a better distribution among responses for each level.

Model Building

When creating a predictive model for whether a LHD has implemented an ICS and/or and EOC, there are multiple options to consider. For this analysis, weighted models were built using SAS 9.3, SUDAAN 10, and WesVarPC 5.1 with binary response data. These models used six variables as possible predictors of whether a LHD

implemented an ICS or an EOC. The variables used were both numerical and categorical, which affects the coding in both SAS and SUDAAN. SAS and SUDAAN were used separately to create predictive models because of the fundamental differences in how each creates variance estimation. SUDAAN gives only logistic regression models divided by levels of each variable. SAS and WesVarPC can give models either divided by level of each variable or not divided. To ensure the reference levels were consistent in all three programs, the reference levels were manually identified in SAS. This was done by adding two options, identifying the reference level after the variable in the model statement and specifying *param=ref*, as seen in Appendix I.

Major differences in the distribution of the responses of variables ICS and EOC exist in the data. Of the 420 observations used for modeling, 12 LHDs reported not having ICS, while 149 LHDs reported not having an EOC. This relatively poor distribution of responses for ICS offers the opportunity to examine how well each statistical package deals with such response variable distributions. While a poor distribution of response variables can lead to a poor predicative model, it must be noted that this analysis does not focus on the predictive value of the models. Instead, this analysis focuses more on the differences in variance estimation produced by each statistical package. Therefore interaction terms were ignored for simplicity. By using data and picking explanatory variables with realistic predictive values, an effort was made to be as realistic as possible.

Logistic Regression

Logistic regression is used in this analysis because the outcome variable is binary in both ICS and EOC models. A simple linear regression (SLR) allows for a direct fitting

of the probability of success. SLR is shown in equation (7) and the basic multiple linear regression (MLR) equation is denoted in (8) where p_i denotes the probability of success.

(7) $p_i = \beta_0 + \beta_1 X,$ predictor variable = X

(8) $p_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$ predictor variables = $X_1, X_2 \ldots X_p$

SLR and MLR are used to build models when the dependent response variable is continuous. When the response variable is binary, using a linear regression is not a good idea since p_i is restricted to [0, 1]. Thus a logistic transformation, as seen in (8), should be performed to change the range from (0, 1) to $(-\infty, \infty)$ (Collett).

(9) $logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$

Using this transformation, the logistic model in (9) is produced and the probability of success is shown in (10).

(10) $logit(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_p$

(11) $p_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_p X_{p_i}}}{1-e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_p X_{p_i}}}$

All three statistical software packaging examined in this analysis use this basic model when creating β and variance estimates.

Taylor Series Linearization

The procedures used in SAS and SUDAAN for this analysis use Taylor Series to create variance estimate. A Taylor Series is used to linearize non-linear functions by creating a polynomial expression. Below is an example of how a function is linearized by Taylor expansion. Imagine that one wants to expand the exponential function below.

(12) $f(x) = e^x,$ center at a = 0

To do this the function is plugged into the following power series known as the Taylor series.

$$(13) \quad \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

The expanded infinite series constructed from this notation is as follows.

$$(14) \quad f(x) = f(a) + f'(a)\frac{(x-a)}{1!} + f''(a)\frac{(x-a)^2}{2!} + f'''(a)\frac{(x-a)^3}{3!} + \cdots$$

Plugging in the exponential function in equation (15) the following Taylor series is created.

$$(15) \quad e^x = e^0 + f'(e^0)\frac{(x-0)}{1!} + f''(e^0)\frac{(x-0)^2}{2!} + f'''(e^0)\frac{(x-0)^3}{3!} + \cdots$$

The value of e^0 is equal to one, and any derivative of one is also one.

$$(16) \quad e^0 = 1,\ f'(1) = 1, f''(1) = 1, \ldots.$$

Plugging these values into equation (17) the final Taylor series is produced.

$$(17) \quad e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,\ -\infty < x < \infty$$

Due to the use of Taylor series linearization, both SUDAAN and SAS need the entire sample and each observation weighted in order to produce the correct variance estimates (Heeringa).

Balanced Repeated Replication

Variance estimation can also be done using replication techniques, such as BRR, jackknife, and bootstrap. BRR is characterized as a "half-sample" method for estimating variance. One requirement of BRR variance estimate is that at most, two PSUs may exist per stratum. It is possible to have a single PSU in a stratum, but only if it indicates a self-representing stratum, i.e. the observation was selected with certainty. This can be specified within the commands of WesVarPC. To produce a replicate, one PSU from

each stratum is selected, forming a "half-sample replicate" as well as its "half-sample complement." This means that by permutation, there are a total of 2^H total possible "half-sample replicates", where H is the number of strata. Not all replicates are needed to obtain the necessary information for variance estimation. There is a minimally sufficient set which if selected, drops out any "unwanted between-stratum cross-product terms" (Heeringa). In WesVarPC this minimally sufficient set is called the balanced half-samples, and is produced by using "orthogonal Hadamard matrices," as seen in Appendix G (Morganstein). The "+" and "-" signs in the matrix indicate which of the PSUs is to be used for the replicate. New weights must be assigned in these replicates since a PSU is dropped from each stratum. These new replicate weights are simply the sample weights multiplied by a value of two.

(18) $w_{i,rep} = w_i * 2$, where w_i = sample weight

The replicate weights are then used to calculate the replicate statistics of interest, using only the selected PSUs, as seen in equation (19) and the overall population statistic (20) using the whole sample.

(19) $$\hat{\theta}_r = \frac{\sum_{i \in rep}^{n_{rep}} \theta_i * w_{i,rep}}{\sum_{i \in rep}^{n_{rep}} w_{i,rep}}$$

(20) $$\hat{\theta} = \frac{\sum_{i=1}^{n} \theta_i * w_i}{\sum_{i=1}^{n} w_i}$$

These values are then plugged into the following BRR equation for variance estimation. For the BRR method, the degrees of freedom are equal to the number of strata.

(21) $var_{BRR}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \hat{\theta})^2$, where replicates = 1,…,R and df = H

These variance estimates are then used to produce confidence intervals over parameters of interest, and are thus reported as standard errors (Heeringa).

Proc Surveylogistic

For the binary response analysis in SAS, a logistic regression model was created using the *proc surveylogistic* function. *Proc surveylogistic* was chosen because the response variable is binary, represented by a value of zero if an ICS did not exist in the LHD, and a value of one if an ICS did exist (in a similar fashion to the EOC variable). No selection method was used for this analysis since the variance estimations were of more interest than the actual validity of the model.

Proc rlogist

For the binary response analysis in SUDAAN, the *proc rlogist* procedure was used to build the models. This procedure was chosen because the response variable is binary, which is similar to the *proc logistic* procedure in SAS, but differs in multiple ways. The response variable can have the zero/one coding, although this is the only time in SUDAAN where zero is an acceptable value. The categorical predictor variables cannot use the value of zero in their coding. Extra statements are added to account for the complex design and the weighting, as well as to produce the variance-covariance matrix for the βs.

WesVar 5.1

Unlike SUDAAN, the WesVarPC statistical software package runs separately from SAS. The dataset was read into WesVarPC from a saved SAS dataset. The stratum and PSU variables were created separately in SAS before being read into WesVarPC. Once read into WesVarPC, a new data file was created which contained the same 355 observations, weights, and variables modeled in SAS and SUDAAN. Once the BRR method was selected, the two self-representing strata were identified so that no error

would occur when run, since there is only one PSU in each of those strata. Once these self-representing strata were identified, new weights were created. A new workbook was used to produce a logistic regression and the variance estimates for the βs. This was done by choosing the logistic option in the regression modeling window.

Results

ICS Model

The results of the analysis, seen in Appendices C and D, have been rounded off to two decimal places. SUDAAN does not give β estimates or standard errors for an overall model in which each variable is not broken down by level. In the overall model, the estimates given by SAS and WesVarPC are the same for all variables and intercept, but WesVarPC gives larger standard error estimates. The standard errors given by WesVarPC are generally much larger than those given by SAS. The smallest difference is seen in the funding variable with a value of 0.27 given by SAS and 0.58 given by WesVarPC. The largest difference can be seen in the intercept with a standard error of 2.44 given by SAS and 34.83 given by WesVarPC.

When the model is broken down into the levels of each variable, the estimates given by all three packages are the same for all variables, except the intercept and the number of responded incidents. SUDAAN consistently produces the lowest standard error for each level of every variable, followed by SAS, and then WesVarPC which gives the largest standard error estimates. The β estimates for the seven levels of the number of incidents differ in all three packages. When there were zero observations in a level of an observed variable, SUDAAN does not give an estimate for these levels, while WesVarPC and SAS give estimates for all levels, as seen in table C-3 in Appendix C.

EOC Model

As in the ICS model, SUDAAN gives only β estimates and standard errors of the variables broken down into levels, not the overall variables. The estimates are the same for the overall model given by WesVarPC and SAS. WesVarPC again consistently gives larger standard errors for all variables.

The β coefficients given by all three statistical software packages are the same for the EOC model when broken down into levels. As in the ICS model, the standard errors are smallest in SUDAAN, followed by SAS, and then WesVarPC.

Discussion

Standard Error Differences

In the overall EOC logistic regression model produced by WesVarPC and SAS, the standard errors are comparable and quite similar, with the estimated standard error in WesVarPC only slightly larger than those predicted by SAS. In contrast, the overall ICS logistic regression model produced in WesVarPC produces much larger standard errors than does SAS. A similar effect is seen in the ICS and EOC regression models when broken down by levels of variables, except the special case of the intercept and number of responded incidents variable discussed later. The WesVarPC standard errors are only slightly larger than the SAS standard errors in the EOC model, while much larger in the ICS model. The SUDAAN standard errors are slightly smaller than the standard errors given by SAS, with the largest difference of 0.02 for the EOC model and 0.04 for the ICS model.

As discussed in the methods section, the EOC response variable can be considered "better" than that of the ICS variable in light of the more evenly distributed responses. Only 12 of 420 LHDs indicated that they did not have an ICS in place, while 149 did not have and EOC. This difference most likely explains larger variation in standard error produced by the three statistical packages for the ICS models. The more evenly distributed response of the EOC variable allows for smaller variance estimation.

Specific Variable Differences

The most striking difference in the β and standard error estimates in this analysis can be seen in table C-3 in Appendix C. This table displays the estimates for the ICS model broken down by level of the variable for the number of incidents with responses

and the intercept. A problem arises, however, since the number of observations in the levels of "2", "4", and "don't know" is zero for those LHDs that do not have ICS, as seen in table A-1 of Appendix A. This difference in β estimates does not arise in the EOC model, due to the fact that no levels of the incidents responded to variable have zero observations, seen in table B-1 of Appendix B. The result of these zero observations differs for each of the three software packages. Since the "don't know" level is used as the reference level, the estimate is zero, as it is for all other reference levels of variables as well. SUDAAN does not produce β estimates for the "2" and "4" levels. Both SAS and WesVarPC produce β estimates for all levels of the variable for number of incidents, but are different for both. The β estimates for the levels of "2" and "4" are considerably closer to zero than the β estimates for the other levels. In regards to the standard error estimates, SUDAAN has the smallest estimates, followed by substantially larger estimates given by SAS, followed by considerably larger estimates with WesVarPC. The intercepts also have different β estimates in all statistical packages, but this is expected since at least one β is not the same as in the other statistical packages.

Conclusions

Deciding which variance estimation to use for a logistic regression of a complex survey depends on three things: the distribution of the response variable, whether levels of a variable have zero observations, and the goal of the analysis. If the response variable has a "poor distribution," with few observations of either yes or no, SUDAAN should be chosen over SAS and WesVarPC for a model broken down by level. If one desires a model not broken down by level, SAS should be chosen over WesVarPC. Even though

SUDAAN and SAS give comparable standard error estimates, SUDAAN's estimates are consistently slightly lower.

The choice of which package should be used is not always clear-cut. For an example, consider the problem presented by the incidents responded to variable in the ICS model. For some levels of the categorical dependent variable there are zero observations (Table A-1 of Appendix A). SUDAAN gives the desired lowest standard errors, but does not give β estimates for those levels lacking observations. SAS and WesVarPC give estimates for all levels of the variable, but SAS has considerably lower standard errors.

In terms of the goal of the analysis, if one wanted only to create a logistic regression model broken down by levels in the variables, SUDAAN is the best choice. However, since SUDAAN does not give β or standard error estimates for an overall model, the use of SUDAAN is limited. To derive both types of models, SAS is the best choice, since it can give both types of model and the standard errors are only slightly larger than those given by SUDAAN.

Bibliography

Agresti A. 2002. *Categorical Data Analysis, Second Edition*. Wiley.

Chen, X. and Gorrell, P. 2004. Variance estimation with complex surveys: some SAS®-SUDAAN comparisons. *Proceedings of NESUG 17*.

Cohen, S. 1997. An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data. *The American Statistician*, **51** (3), 285-292.

Collett D. 2003. *Modeling Binary Data. Second Edition*, Chapman & Hall.

Herringa, S.G., West, B.T., and Berglund, P.A. 2010. *Applied Survey Data Analysis*, Taylor & Francis, Boca Raton.

Koh, H.K., Elqura, L.J., Judge, C.M., and Stoto, M.A. 2008. Regionalization of Local Public Health Systems in the Era of Preparedness. *Annual Review of Public Health*, **29**, 205-218.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. 2005. *Applied Linear Statistical Models*, McGraw-Hill, New York.

Laurie, N., Wasserman, J., and Nelson, C.D. 2006. Public Health Preparedness: Evolution or Revolution? *Health Affairs*, **4**, 933-945.

Lu, M. and Yang, W. 2012. Multivariate Logistic Regression Analysis of Complex Survey Data with Application to BRFSS Data. *Journal of Data Science*, **10**, 157-173.

Morganstein, D. 1998. *The replication method for estimating sampling errors*, Eustat.

National Research Council. "Front Matter." *The Future of Public Health*. Washington, DC: The National Academies Press, 1988.

Park, M.Y. 2008. Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**(1), 30-50.

Roberts, G. Rao, J.K., and Kumar, S. 1987. Logistic Regression Analysis of Sample Survey Data. *Biomtrika*, **74**(1), 1-12.

Scott, C. 2010. Lessons Learned for H1N1-Use of the Incident Command System in the Public Health Response, *NACCHO*.

Westat, 2007. *WesVar® 4.3 User's Guide*, Westat.

Wolter, K. 1985. *Introduction to Variance Estimation*, Springer-Verlag, New York.

References

California Department of Food and Agriculture[Internet]. State of California. 2012. Available from: http://www.cdfa.ca.gov/ahfss/emergency_preparedness/ICS.html.

Cox, D. 2011. Evolution of the Public Health EOC: One Local Public Health Official's Journey. [PowerPoint slides]. Presented at Heartland Center for Public Health Preparedness Webinar Series.

FSRC Food Safety Information Infrastructure Project: Phase 1. Workshop on Public Sector Food Safety Data Collection, Access and Sharing. *Difference in State and Local Organization: A Picture of Complexity*. November 2-3, 2006.

Hecht, R. L. and Kaufman S. G. 2005. Incident Command System & Public Health: S.Y.S.T.E.M.S. Training. [PowerPoint Slides]. Rollins School of Public Health, Center for Preparedness and Research.

Irwin, R. 2009. Chapter 7: The Incident Command System (ICS). *Disaster Response: Principals of Preparation and Coordination*.

Mindlin, M., and Bilter, E. 2011, Survey Implementation Process v5.

North American Stata Users Group. *Computing Variances from Data with Complex Sampling Designs: A Comparison of Stata and SPSS*, by Dowd, A. 2000.

SAS Institute Inc. *Performing Logistic Regression in Survey Data with the New SURVEYLOGISTIC Procedure*, Working Paper No. 258-27, by An, A. B. Carey (NC).

U.S. Department of Education. National Center for Education Statistics. *Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances*.

from NCES Data Sets, Working Paper No. 2000-03, by Pam Broene and Keith Rust. Project Officer, Susan Ahmed. Washington, DC: 2000.

U.S. Department of Health & Human Services [Internet]. 2009. Available from: http://www.hhs.gov/disasters/discussion/planners/mscc/appendix/b.html.

Wikipedia contributors. Incident Command System [Internet]. Wikipedia, The Free Encyclopedia; 2012. Available from: http://en.wikipedia.org/wiki/Incident_Command_System.

Appendix A: ICS Tables of Distribution for Variables

A-1: ICS Distribution of Number of Incidents Responded To

		Number of Incidents Responded To						
ICS Implemented		**0**	**1**	**2**	**3**	**4**	**5 or more**	**Don't Know**
	No	9	1	0	1	0	1	0
	Yes	128	68	60	43	16	71	20

A-2: ICS Distribution of Number of Workers

		Number of Workers		
ICS Implemented		**> 9**	**10-49**	**≤ 50**
	No	8	3	1
	Yes	100	16	140

A-3: ICS Distribution of Area Served

		Area Served			
ICS Implemented		**Urban**	**Suburban**	**Town**	**Rural**
	No	1	1	2	8
	Yes	59	55	44	247

A-4: ICS Distribution of Part of State Agency

		Part of State Agency	
ICS Implemented		**Yes**	**No**
	No	8	4
	Yes	196	210

A-5: ICS Distribution of Unit for Emergency Preparedness

		Unit for Emergency Preparedness	
ICS Implemented		**Yes**	**No**
	No	2	10
	Yes	244	161

A-6: ICS Distribution of Funding

		Funding				
ICS Implemented		**None**	**$1-24,999**	**$25,000-99,999**	**$100,000+**	**Don't know**
	No	3	5	2	1	1
	Yes	14	71	127	125	69

Appendix B: EOC Tables of Distribution for Variables

B-1: EOC Distribution Number of Incidents Responded To

	Number of Incidents Responded To							
EOC Implemented		**0**	**1**	**2**	**3**	**4**	**5 or more**	**Don't Know**
	No	62	36	23	10	2	12	4
	Yes	75	33	37	34	14	60	16

B-2: EOC Distribution of Number of Workers

	Number of Workers			
EOC Implemented		**> 9**	**10-49**	**≤ 50**
	No	60	60	28
	Yes	48	109	113

B-3: EOC Distribution of Area Served

	Area Served				
EOC Implemented		**Urban**	**Suburban**	**Town**	**Rural**
	No	16	13	21	98
	Yes	44	43	25	157

B-4: EOC Distribution of Part of State Agency

	Part of State Agency		
EOC Implemented		**Yes**	**No**
	No	71	77
	Yes	133	137

B-5: EOC Distribution of Unit for Emergency Preparedness

	Unit for Emergency Preparedness		
EOC Implemented		**Yes**	**No**
	No	63	83
	Yes	183	88

B-6: EOC Distribution of Funding

		Funding				
EOC Implemented		**None**	**$1-24,999**	**$25,000-99,999**	**$100,000+**	**Don't know**
	No	15	30	53	22	29
	Yes	2	46	76	104	41

Appendix C: ICS β Estimates and Standard Errors

C-1: Overall ICS Model

Variable	β Estimate	WesVar	SUDAAN	SAS
		Std. Error	Std. Error	Std. Error
Intercept	-2.03	34.83	----------	2.44
Area served	-0.47	4.12	----------	0.30
Part of State Health Agency	0.13	6.40	----------	0.84
Unit dedicated to Emergency Preparedness	2.09	6.92	----------	0.92
Number of workers	-0.37	10.08	----------	0.64
Number of incidents responded to	-0.32	6.65	----------	0.27
Funding	-0.62	0.58	----------	0.27

C-2: ICS Model by Level, Variable with same β estimates

Variable	β Estimate	WesVarPC	SUDAAN	SAS
		Std. Error	Std. Error	Std. Error
Area served				
Urban	1.66	19.61	1.64	1.67
Suburban	1.97	22.19	1.84	1.88
Town	2.27	15.48	1.01	1.03
Rural	reference			
Part of State Health Agency				
Yes	1.64	12.72	0.93	0.95
No	reference			
Unit dedicated to Emergency Preparedness				
Yes	-3.15	16.1	1.1	1.12
No	reference			
Number of workers				
Fewer than 9	2.05	22.05	1.74	1.78
10-49	1.39	17.26	1.31	1.34
50 or more	reference			
Funding				
None	1.46	15.85	1.51	1.54
$1-24,999	4.1	17.17	1.39	1.42
$25,000-99,999	2.1	16.91	1.51	1.54
$100,000+	-0.11	22.04	2.45	2.51
Don't know	reference			

C-3: ICS Model by Level, Variables with different β estimates

Variable	WesVarPC		SUDAAN		SAS	
	β Estimate	Std. Error	β Estimate	Std. Error	β Estimate	Std. Error
Intercept	21.42	42.45	16.26	2.46	24.71	4.73
Number of incidents						
0	-13.90	10.35	-8.74	0.82	-17.19	3.69
1	-12.74	21.75	-7.58	1.33	-16.03	3.83
2	2.31	16.20	----------	0.85	1.83	3.76
3	-15.21	17.75	-10.05	0.90	-18.5	3.76
4	2.00	22.09	----------	0.93	1.99	4.06
5 or more	-10.91	23.39	-5.75	2.65	-14.20	4.26
Don't know	reference		reference		reference	

Appendix D: EOC β Estimates and Standard Errors

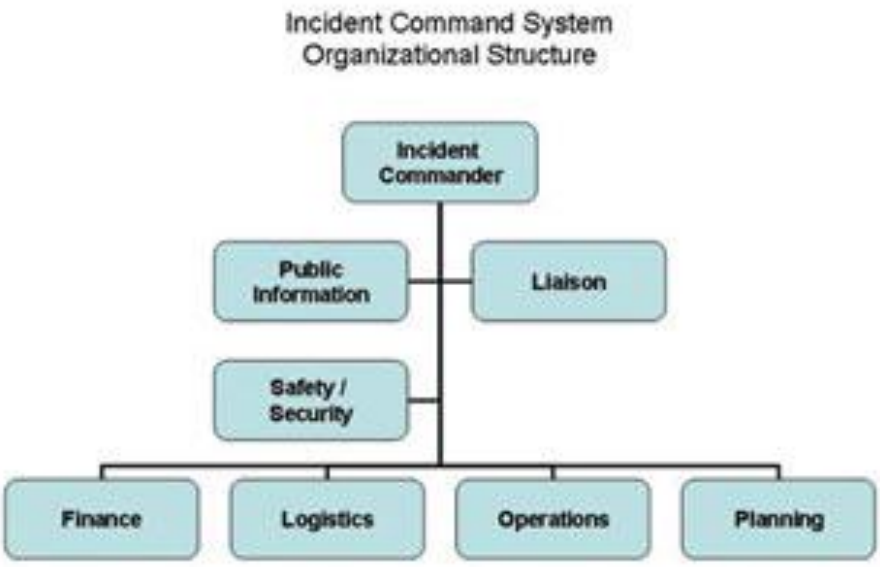
D-1: Overall EOC Model

Variable	B Estimate	WesVarPC	SUDAAN	SAS
		Std. Error	Std. Error	Std. Error
Intercept	-0.76	1.67	----------	1.36
Area served	0.06	0.19	----------	0.15
Part of State Health Agency	-0.29	0.42	----------	0.35
Unit dedicated to Emergency Preparedness	-0.28	0.41	----------	0.35
Number of workers	0.85	0.31	----------	0.26
Number of incidents responded to	0.09	0.1	----------	0.08
Funding	-0.02	0.22	----------	0.19

D-2: EOC Model by Level

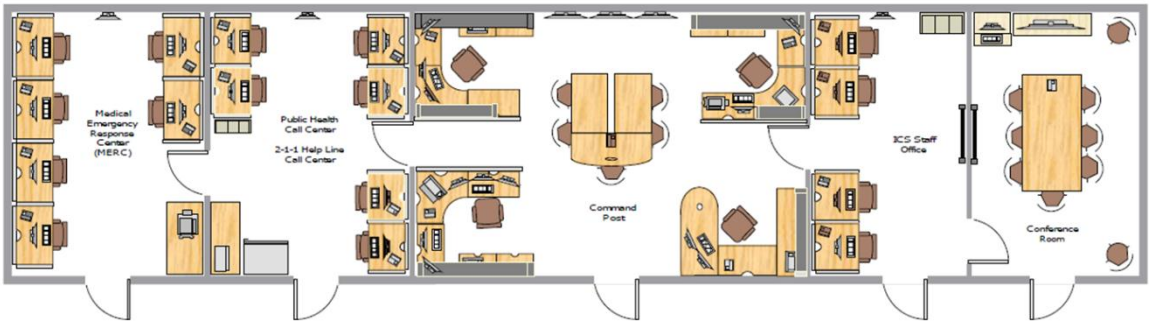
Variable	β Estimate	WesVarPC	SUDAAN	SAS
		Std. Error	Std. Error	Std. Error
Area served				
Urban	-0.39	0.65	0.46	0.47
Suburban	-0.50	0.64	0.45	0.46
Town	-0.73	0.75	0.56	0.57
Rural	reference			
Part of State Health Agency				
Yes	0.37	0.40	0.31	0.31
No	reference			
Unit dedicated to Emergency Preparedness				
Yes	0.26	0.36	0.30	0.31
No	reference			
Number of workers				
Fewer than 9	-1.10	0.57	0.43	0.43
10-49	-0.43	0.44	0.36	0.37
50 or more	reference			
Funding				
None	-0.01	8.80	1.07	1.09
$1-24,999	0.13	0.67	0.51	0.53
$25,000-99,999	0.46	0.57	0.42	0.42
$100,000+	1.43	0.65	0.48	0.49
Don't know	reference			
Number of incidents responded to				
0	-1.35	3.23	0.88	0.90
1	-2.62	3.21	0.88	0.90
2	-1.75	3.23	0.86	0.88
3	-1.61	3.21	0.90	0.91
4	-1.17	7.67	1.09	1.12
5 or more	-1.49	3.19	0.86	0.87
Don't know	reference			

Appendix E: Basic ICS Organizational Chart



(California Department of Food and Agriculture, 2012)

Appendix F: Integrated EOC Model



(Cox, 2011)

Appendix G: Orthogonal Hadamard Matrix where H=4

BRR Replicate	Stratum (h)			
	1	2	3	4
1	+	+	+	-
2	+	-	-	-
3	-	-	+	-
4	-	+	-	-

(Heeringa, 2010)

Appendix H: SAS Code: Program 1

```
***Sarah Robertson
Thesis ICS/EOC Predictive Modeling
Modeling Dataset Creation***
;

*create strata to select by;

data work.newskipslb;
      set perrc.newskipslb;

      **create new variable;
      if ics = 1 and eoc = 1 then strata=1; else
      if ics = 1 and eoc = 2 then strata=2; else
      if ics = 2 and eoc = 1 then strata=3; else
      if ics = 2 and eoc = 2 then strata=4; else
      strata = 5;

      *create variable for stratum;
      if stratum = 301 then strat=1; else
      if stratum = 104 then strat=1; else
      strat=2;

      keep  ics eoc strata stratum workers area
                  statepart statediff staterequire
                  unit funding resp participate whenh1n1
                  leaderfeel nonepfeel finalwt strat pop stratx k;
run;

*Recode variables;


data work.newskipslb;
      set work.newskipslb;

      *only keep if response variables are yes or no;
      if strata lt 5;

      *recode ICS and EOC variables for SUDAAN
            0 = "No"
            1 = "Yes";
      if ics lt 3;
      if eoc lt 3;
      if eoc = 2 then eoc = 0;
      if ics = 2 then ics = 0;

      *recode funding variable;
      if funding = 1 then fund=1;else
      if funding = 2 then fund = 2; else
      if funding = 3 then fund = 2; else
      if funding = 4 then fund = 3; else
      if funding = 5 then fund = 3; else
      if funding = 6 then fund = 4; else
      if funding = 7 then fund = 4; else
      if funding = 8 then fund = 5;
```

```
      *recode workers variable;
      if workers = 1 then work = 1; else
      if workers = 2 then work = 1; else
      if workers = 3 then work = 2; else
      if workers = 4 then work = 2; else
      if workers = 5 then work = 3; else
      if workers = 6 then work = 3; else
      if workers = 7 then work = 4;

      if unit le 2;

run;

*set up psu and for modeling by strat and size;

*sort by size;

proc sort data=work.newskipslb;
      by strat pop;
run;

*create psu for SAS and SUDAAN;
*create _psu for WesVarPC. Needs to be coded using
the value of 1 and 2 for _psu;

data work.newskipslb2;
      set work.newskipslb;

      obs + 1;

      *create group for strat = 2;
      i = 420;
      n = 211;

      do until (i= 2);

            if obs le i then group = n;
            i = i - 2;
            n = n - 1;
      end;

      *create psu for strat = 1;
      if obs le 2 then group = obs;

      *create  psu;
       psu = 2;
      k = 1;

      do until (k=421);

            if obs = k then _psu = 1;

            k = k + 2;
      end;

run;
```

```
*check;

proc print data=work.newskipslb2;
      var strat pop obs group _psu;
run;

*create permanent datasets;

data home.modeling;
      set work.newskipslb2;

      keep  ics eoc
                  work area statepart unit fund resp
                  finalwt
                  strata strat stratum stratx
                  group _psu obs;
run;
```

Appendix I: SAS Code: Program 2

```
***Sarah Robertson
Thesis ICS/EOC
Predictive Modeling***
;


*** SAS Predictive Models

***1. Univariate SAS ICS Model;

proc surveylogistic data=home.modeling;

      strata stratx;

      model ics (event="1") = area statepart unit work resp fund ;

      weight finalwt;

run;

proc surveylogistic data=home.modeling;

      strata stratx;

      class       area(ref='4')statepart unit(ref='2')work(ref='3')
                  resp(ref='7')fund(ref='5')/param=ref order=intenal ;


      model ics (event="1") = area statepart unit work resp fund ;

      weight finalwt;

run;

***2. Univariate SAS EOC Model***;

proc surveylogistic data=home.modeling;

      strata stratx;

      model eoc (event="1") =  area statepart unit work resp fund;

      weight finalwt;

run;

proc surveylogistic data=home.modeling;

      strata stratx;

      class       area(ref='4')statepart unit(ref='2')work(ref='3')
                  resp(ref='7')fund(ref='5')/param=ref order=intenal ;

      model eoc (event="1") =  area statepart unit work resp fund;
```

```
      weight finalwt;

run;


***SUDAAN  Models;

*** 1. ICS model;

proc sort data=home.modeling;
      by stratx ;
run;

proc rlogist data=home.modeling design=wr filetype=SAS ;

      nest stratx ;
      weight finalwt;

      setenv colwidth=20;

      subgroup area statepart unit work resp fund ;
      levels     4      2     2     3     7   5 ;


      model ics= area statepart unit work resp fund ;



run;

***2.  EOC model;

proc rlogist data=home.modeling design=wr filetype=SAS ;

      nest stratx;
      weight finalwt;


      subgroup area statepart unit work resp fund;
      levels      4     2     2     3     7   5;

      model eoc = area statepart unit work resp fund;

run;
```