THE ROLE OF WEIGHTS IN MULTIVARIATE ANALYSES OF THE NCVS

Sharon Lohr, Arizona State University; Joanna Liu, Unisys Corporation
Sharon Lohr, Dept. of Mathematics, Arizona State University, Tempe, AZ 85287-1804

KEY WORDS: sampling weights, logistic regression

Design and Weighting of the NCVS. The National Crime Victimization Survey (NCVS), administered by the Bureau of Justice Statistics, is a survey of persons across the U.S. in which noninstitutionalized persons 12 years old and older are asked about their experiences as victims of crime in the past six months. The NCVS takes a stratified, multistage cluster sample that is designed to be approximately self-weighted.

Even though the NCVS is designed to be self-weighted, weights are used to adjust for some differential sampling rates, to adjust for nonresponse, and to adjust for differences between the demographic composition of the sample and that of the U.S. noninstitutionalized population aged 12 and over. The design of the NCVS and the procedure followed for calculating weights are described in U.S. Department of Justice (1991, 1992). The weight variable for a person in the sample is intended to be the number of persons in the target population represented by that person.

In this paper we discuss rationales for and against using weights in linear and logistic regression. We fit a number of logistic regression models similar to those that might be used in more detailed studies concerning correlates or dynamics of victimization to see how much difference the use of weights makes. Finally, we offer some advice for deciding whether or not to use weights in an analysis.

Weighted vs. Unweighted Analyses. There is little argument that weights should be used in calculating means and totals, as in the annual reports of *Criminal Victimization in the United States* (U.S. Department of Justice, 1992). This has been justified from both design-based and model-based perspectives, as in Little (1991) and Kish (1992).

There has been much vigorous debate, however, on whether weights should be used in more complex analyses. Much work done on the effects of weights in complex surveys has focused on contingency table analysis or on linear regression. In linear regression, the debate between weighted and unweighted analyses can be summarized as follows. Let Y be the vector of responses, and X the matrix of explanatory variables for the whole population. Then

$$B = (X'X)^{-1}(X'Y) \tag{1}$$

is the population regression parameter. In practice, only a sample of observations is taken, with observation vector y and matrix of explanatory variables x. This sample must be used to obtain summary statistics relating Y and X.

A person holding a design-based perspective takes the point of view that the population values Y are fixed: a probability structure only occurs because some units were randomly chosen to be included in the sample while others were not. The goal is to estimate B, regardless of population structure. Since B can be written as a function of population means and totals, one should (by analogy with the practice for estimating means) obtain weighted estimates of those means and totals to give

$$b_w = (x'wx)^{-1}x'wy \tag{2}$$

where w is the diagonal matrix of weights.

A person holding the model-based perspective instead postulates that a stochastic model describes the relation between y_i and x_i:

$$y_i = x_i^T\beta + \epsilon_i. \tag{3}$$

If all observations in the population follow the model (3), then the sampling design should have no effect as long as the sampling is not outcome-based. The value B is the least squares estimate of β if the whole population were observed: since only a sample is observed, one should use the least squares estimate

$$b = (x'x)^{-1}x'y. \tag{4}$$

One searches for a structure that can be thought to generate the population and then estimates the parameters for that structure.

An advocate of b_w is often concerned about possible biases arising from the sampling design. Kott (1991) and authors referenced in Kott's paper compare design-based and model-based estimates in linear regression. He argues that sampling weights are needed in linear regression because the choice of covariates in survey data is limited to variables collected in the survey: if necessary covariates are

omitted, b_w and b are both biased estimators of β, but the bias of b_w is a decreasing function of the sample size, while b is only asymptotically unbiased if the probabilities of selection are not related to the missing covariates.

The robustness to missing covariates, however, comes at a cost. The standard errors of the weighted parameter estimates are generally higher than the corresponding unweighted estimates, as discussed in Kish (1992).

Fienberg (1980) says "we know of no justification whatsoever for applying standard multivariate methods to weighted data . . . the automatic insertion of a matrix of sample-based weights into a weighted least-squares analysis is more often than not misleading, and possibly even incorrect." Saphire (1984, p. 38) also takes a model-based approach to estimation in the NCVS.

Smith (1988) and Little (1991) adopt a model-based perspective but argue that sampling weights are useful in model-based inference as summaries of covariates describing the mechanism by which units are included in the sample. Their argument can be extended to justify weights to adjust for undercoverage and nonresponse by using weights as stand-ins for variables affecting nonresponse. In the NCVS, modeling assumptions are used in constructing those weights, and several variables used for constructing the weights used to adjust for nonresponse are available for use as covariates in models.

Weights in Logistic Regression. In logistic regression the response is binary rather than continuous, taking on the value of 1 or 0. Let x be a p-vector of independent variables, and β be the p-vector of unknown parameters.

Much of the above discussion on linear regression also applies to logistic regression. Binder (1983) gives design-based theory for estimating logistic regression parameters. He starts with a parametric likelihood function for logistic regression, the likelihood that would be adopted if the entire population were available for study and if the observations were independent. The finite population parameter B is then defined to be the maximum likelihood estimate of β. If all N elements in the population could be observed, the parameter B would be the solution to the system of equations

$$\sum_{i=1}^{N} x_{ij}(y_i - \frac{\exp(x_i^T B)}{1+\exp(x_i^T B)}) = 0 \qquad (5)$$

for $j = 1, \ldots, p$.

Although B is defined in terms of the likelihood, Binder (1983) proposes estimating B rather than β. An estimate of B is given by the solution b_w to

$$\sum_{i \in \mathcal{S}} w_i x_{ij}(y_i - \frac{\exp(x_i^T b_w)}{1+\exp(x_i^T b_w)}) = 0 \qquad (6)$$

for $j = 1, \ldots, p$, where $\mathcal{S}$ denotes the units included in the sample. The i^{th} observation in the sample represents w_i observations in the population.

For the data in the sample, the quasi-likelihood estimates b_u solve equation (6) when the w_i are set equal to one. Even though the covariance matrix is misspecified, b_u is still a consistent and asymptotically normal estimator for β if the model is correct.

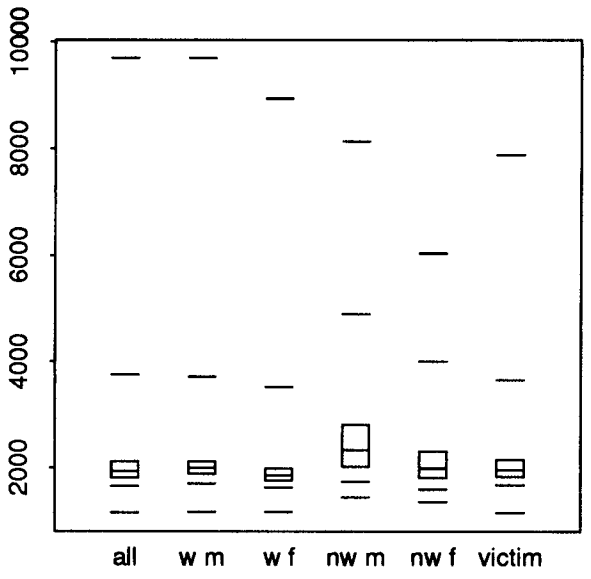
Under the assumption that the logistic model is correct and that the marginal probability distribution of x contains no information about the coefficients in the logistic regression, Prentice and Pyke (1979) showed that case-control models may be analyzed as cohort studies in logistic regression: the only parameter estimate that is affected by sampling dependent on the y's is the intercept. Nonresponse can be considered as a modified version of a case-control study: the units selected to be in the sample are in two strata, respondents and nonrespondents. Thus we might expect that because of this special property of logistic regression, using sampling weights would have little effect on coefficients other than the intercept if all important independent variables are included in the model.

Scott and Wild (1989) find that the Prentice and Pyke results hold for stratified sampling when the strata are determined by the value of the binary response variable. They conclude that when the model fits well, there is little difference between model-based and design-based point estimates for the parameters, but that the model-based estimates are more efficient. When covariates are missing, however, and an estimate of B is desired, they show that the bias of b_u is greater than the bias of b_w.

An argument similar to that in Little (1991) shows that the weighted estimates will correspond to the posterior distribution of the parameters when the coefficients are assumed to be different in each cluster; the unweighted estimates will correspond to the posterior distribution when all clusters are assumed to follow the same model.

The above discussion applies only to point estimates of the parameters. Whether weighted or unweighted estimates are used, the sampling design of a complex survey such as the NCVS should always be taken into account when estimating variances of the parameter estimates. If the dependence among

Figure 1: Boxplots of the distribution of weights for all data, white males, white females, nonwhite males, nonwhite females, and violent crime victims. The horizontal lines from top to bottom represent the maximum, 95th percentile, 75th percentile, median, 25th percentile, 5th percentile, and minimum value in the groups.



observations is ignored, estimates of variances will generally be too small.

The Data Set and Analysis. The data tape used was the NCVS 1990 person-level file (U.S. Department of Justice, 1991). All persons who were interviewed between July and December, 1990 were included in the data set for studying victimization. If a person was victimized more than once, only one incident (chosen randomly) was used in any given model. This was intended to reduce the clustering effect of having more than one crime per person. Weights for nonvictims were multiplied by 10 since the file included only a 10% sample of nonvictims.

Secondary data analysts of the NCVS cannot calculate variances of the regression parameters because clustering information is not released. Even when non-weighted estimates are used, variance estimates given by standard statistical packages will be wrong because observations are dependent. We can conjecture, following Skinner (1989), that the design effect for parameter estimates is larger than one, and that the standard errors given by standard logistic regression procedures are, if anything, too small.

We looked at progressively smaller subsets of the sample in constructing logistic regression models, examining correlates of victimization, reporting crime to the police, and injury. Only the analyses for violent crime victimization and reporting violent crime to the police are presented here, as they show the greatest difference between the weighted and unweighted analyses; other analyses are reported in Liu (1992). In the data set, there were 1092 violent crimes; of the victims of violent crimes, 499 reported the crime to the police.

Box plots for the distribution of weights for the data set in this paper are shown in Figure 1. The distributions of weights are skewed, but most weights in the data set fall within a narrow range. The distribution of weights is fairly homogeneous for various subsets of the data; the largest differences occur among the different race/gender groups, where nonwhite males tend to have higher weights.

Models for Violent Crimes. *Criminal Victimization in the United States, 1990* (U.S. Department of Justice, 1992) gives information on who was most likely to be victimized by various types of crime in 1990. The most likely persons to be victims of violent crime were under 25, black, male, never married, from low-income families, and residents of central cities, among other factors. These variables and their interactions were used as a starting point for building the logistic regression models. Only demographic variables readily obtained from the NCVS were used in this paper; a person's routine activities and lifestyle also affect the likelihood of violent victimization, but the NCVS gives only crude measures of these variables.

Most of the independent variables used in this paper are dichotomous, taking on the value 0 or 1. The "1" values of the variables are:

race	nonwhite
sex	female
age	age 25 or over
married	1 = married or widowed
move5	moved at least once in last five years
offwepn	offender had a weapon
centcity	live in central city of MSA
anyinj	someone injured in incident
rent	housing unit is rented

The variable "income" is retained as an ordinal variable, with the lowest value 1 corresponding to a family income of less than $5000 and the highest value 14 corresponding to a family income of more than $75000. Plots of the data revealed a straight-line relationship between the variable income and

Table 1: Estimated coefficients for a multiple logistic regression model for the response VIOLENT. VIOLENT = 1 if person was victim of violent crime, 0 if person not victim of crime.

Variable	Coef. (w/o wgts.)	Coef. (with wgts.)	Std. error
intercept	-1.32	-3.66	.120
sex	-0.50	-0.49	.070
age	-0.51	-0.46	.078
race	-0.24	-0.32	.099
married	-1.12	-1.16	.085
move5	0.75	0.76	.080
rent	0.32	0.29	.084
centcity	0.36	0.36	.074
income	-.024	-0.023	.009

the various responses in the logit scale.

Table 1 gives parameter estimates for logistic regressions using the response VIOLENT of violent victimization when variables used in constructing weights (for example sex, race, and age) were included as covariates. All standard errors given are those from the unweighted analysis, and must be taken as underestimates of the error. Even with the underestimated standard errors, though, it is seen that, except for the intercepts, the parameter estimates do not change by a large amount when an analysis with weights is performed. A model fit without the variables sex, race, and age shows no difference between the weighted and unweighted estimates.

A positive coefficient indicates that, after controlling for all other independent variables in the model, a higher value of the explanatory variable is associated with a higher likelihood of being a victim of violent crime. Of course, one must always be careful in interpreting such models because of the high degree of multicollinearity in the explanatory variables. The variables rent, centcity, income, and race are strongly associated and must be interpreted together. This association is seen when race is removed from the model in Table 1: the coefficients of rent and income both change.

Note that the largest change in coefficients in Table 1 is for the intercept, as would be expected from the results in Prentice and Pyke (1979), since we multiplied the weights of all nonvictims by ten. The next largest differences are for the variables race and age, both of which are variables used in constructing the weights. It is not surprising that the coefficient of race changes significantly in the weighted analysis for this example, as Figure 1 showed that the weights for nonwhite males were more variable than the weights for other race/gender groups.

Table 2: Estimated coefficients for a multiple logistic regression model for the response RPTPOL, among victims of violent crime.

Variable	Coef. (w/o wgts.)	Coef. (with wgts.)	Std. error
intercept	-0.88	-0.90	.188
sex	0.46	0.45	.143
race	-0.68	-.57	.266
race*sex	0.99	0.79	.376
anyinj	0.70	0.77	.132
age	0.42	0.34	.143
offwepn	0.63	0.62	.140
income	-0.041	-0.040	.017
married	0.41	0.59	.161

The difference in the weighted and unweighted estimates in Table 1 may suggest that some necessary covariates are missing. When the same model used in Table 1 is fit only to the nonwhite persons in the sample, the coefficients for move5, centcity, and age change dramatically, indicating the presence of interactions between race and these variables. Thus the difference in the weighted and unweighted estimates can provide a clue that different regression models may be needed for subclasses of the data.

A model for predicting RPTPOL for victims of violent crime (where RPTPOL = 1 if someone reported the crime to the police, and 0 otherwise) is given in Table 2. Positive coefficients are associated with higher rates of reporting crime to the police. Again, models were basically unchanged with the addition of weights, except for the factors involving race and age.

Conclusions. In all models we fit, both those reported here and those in Liu (1992), few major conclusions about the data would be affected by a decision to include rather than exclude the weights in the analysis. No variables changed sign as a result of including weights, and changes in the coefficients were generally less than one (unweighted analysis) standard error, except in the case of the intercept. The models in this paper represent the largest changes in the models that we found; for most models we fit, the coefficients were the same to two decimal places for the weighted and unweighted

analyses.

Primary purposes of the weights in the NCVS are to compensate for undercoverage and nonresponse, and to adjust for sampling variability by adjusting estimated population totals from the sample so that they agree with independent estimates of population totals. As Kish (1992, p. 186) says: "When nonresponses are not high, the differences between subclasses tend to be small, and then small differences in weights will not have large effects on combined results." Kish's statement is consistent with the results we found for the NCVS, in which the nonresponse rate tends to be around 5%.

Because the NCVS is designed to be self-weighting, most units do not have large differences in the probabilities of selection. Thus our result that use of weights does not have much impact on the conclusions from the data analysis does not generalize to all other surveys. Not all surveys follow a self-weighting design: for surveys in which strata are sampled at different rates, the probabilities of including units in the sample will vary from stratum to stratum, and it is very possible that weighted models would lead to different conclusions than unweighted models. Such a situation would indicate that different models would be appropriate in different strata, and while our preference would be to investigate the underlying mechanisms rather than automatically using weights, Little (1991) and Smith (1988) argue that sampling weights can serve as a proxy for missing stratum information.

We view the use of weighted models as a tool in model development: a practically important difference in the coefficients between weighted and unweighted models indicates a need for further statistical investigation. Strong assumptions about the data are made when the weight variables are constructed for the NCVS data. One assumption made is that nonresponding persons within a weighting class can be represented by the responding persons within that weighting class. These assumptions are (or should be!) acknowledged explicitly when the NCVS is used to estimate victimization rates; they are not often (but again should be) acknowledged when methodological models are fit using the weights. If possible, we believe it is preferable to explicitly account for the nonresponse assumptions and missing covariates directly in the model rather than to use weighting variables. If this cannot be done, though, and weights are used, we believe that the assumptions underlying the weights should be emphasized in the analysis, and the researcher should state that the weights may be absorbing some possible interactions in the model.

References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.*, 51: 279-292.

Fienberg, S.E. (1980). The measurement of crime victimization: prospects for panel analysis of a panel survey. *The Statistician*, 29: 313-350.

Kish, L. (1992). Weighting for Unequal P_i. *J. Official Statistics*, 8: 183-200.

Kott, P.S. (1991). A model-based look at linear regression with survey data. *The American Statistician*, 45: 107-112.

Little, R.J.A. (1991). Inference with survey weights. *J. Official Statistics*, 7: 405-424.

Liu, J. (1992). The effects of weights in logistic regression analyses of National Crime Victimization Survey near-term data. Unpublished M.S. thesis, Arizona State University.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66: 403-411.

Saphire, D.G. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey.* Lecture Notes in Statistics Vol. 23. New York: Springer-Verlag.

Scott, A.J. and Wild, C.J. (1989). Selection based on the response variable in logistic regression. In Skinner, C.J., Holt, D., and Smith, T.M.F. (eds.), *Analysis of Complex Surveys.* New York: Wiley, 191-205.

Skinner, C.J. (1989). Domain means, regression, and multivariate analysis. In Skinner, C.J., Holt, D., and Smith, T.M.F. (eds.), *Analysis of Complex Surveys.* New York: Wiley, 59-87.

Smith, T.M.F. (1988). To weight or not to weight, that is the question. In Bernardo, J.M., DeGroot, M.H., Lindley, D.V., and Smith, A.F.M. (eds.), *Bayesian Statistics 3.* London: Oxford University Press, 437-451.

U.S. Department of Justice, Bureau of Justice Statistics (1991). *National Crime Surveys: National Sample, 1986-1989 [Near-term data]* [Computer file]. 3rd ICPSR ed. Conducted by the U.S. Dept. of Commerce, Bureau of the Census. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor].

U.S. Department of Justice, Bureau of Justice Statistics (1992). *Criminal Victimization in the United States, 1990.* Washington, D.C.: U.S. Government Printing Office.