Most major population surveys used by social scientists are based on complex sampling designs where sampling units have different probabilities of being selected. Although sampling weights must generally be used to derive unbiased estimates of univariate population characteristics, the decision about their use in regression analysis is more complicated. Where sampling weights are solely a function of independent variables included in the model, unweighted OLS estimates are preferred because they are unbiased, consistent, and have smaller standard errors than weighted OLS estimates. Where sampling weights are a function of the dependent variable (and thus of the error term), we recommend first attempting to respecify the model so that they are solely a function of the independent variables. If this can be accomplished, then unweighted OLS is again preferred. If the model cannot be respecified, then estimation of the model using sampling weights may be appropriate. In this case, however, the formula used by most computer programs for calculating standard errors will be incorrect. We recommend using the White heteroskedastic consistent estimator for the standard errors.

Sampling Weights and Regression Analysis

CHRISTOPHER WINSHIP
Harvard University
LARRY RADBILL
Joint Economic Committee of Congress

INTRODUCTION

Most of the major population surveys used by social scientists are not created using the simple random samples discussed in statistics texts. Instead, they are based on complex sample designs where sampling units (e.g., individuals) frequently have different probabilities of being selected. In order to use data of this type to produce descriptive (univariate) estimates of the population, sampling weights (sometimes also called population weights) must be used. It is less obvious, however, whether these sampling weights should be used when estimating regression equations.

This article is aimed at the empirical practitioner who has a good understanding of basic regression analysis. We hope to provide some practical guidance for dealing with sampling weights in regression analysis that will apply in a wide variety of situations.[1] We do not discuss the use of sampling weights in other contexts or the use of other types of weights in regression analysis. In the interest of accessibility, we have attempted to minimize notational complexity by avoiding matrix notation. All the results in the article can be readily generalized to situations where there are any number of independent variables. The appendix provides formulas using matrix notation.

We argue that there are pitfalls to the common practice of estimating ordinary least squares (OLS) models with weighted data. The most troublesome problem is that almost all computer packages use the incorrect formula to estimate coefficient standard errors when sampling weights are used. This has potentially important implications for decisions about model specification and formal hypothesis testing that may be based on miscalculated numbers.

In deciding whether to use weighted data in a regression analysis, we distinguish between two situations: where the weights are solely a function of observed independent variables included in the model and where the weights are also a function of the dependent variable and thus the error term. In the first situation, use of the sampling weights (weighted ordinary least squares, or WOLS) yields unbiased and consistent parameter estimates, but OLS also provides unbiased and consistent estimates with smaller standard errors. (See appendix for formulas.) Here unweighted OLS is the preferred approach. In the second case, we recommend first attempting to respecify the model so that the weights are solely a function of the independent variables. If this can be accomplished, then unweighted OLS will yield unbiased, consistent, and efficient parameter estimates. Weighted estimates will also be unbiased and consistent, but not efficient. If the model cannot be respecified, then estimation of the model using WOLS may be appropriate. However, if WOLS is used, the formula used by most computer programs for calculating standard errors of the estimates will be incorrect and should not be used. When WOLS is used, we

AUTHORS' NOTE: *The authors have benefited greatly from comments provided by Christopher Jencks, Ian Domowitz, Bruce Spencer, Scott Long, and anonymous reviewers. Of course, any remaining mistakes in the article are alone our responsibility. The views expressed in this article are those of the authors. They do not necessarily represent the views of the Joint Economic Committee or any of its members.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 23 No. 2, November 1994 230-257

recommend using the White (1980) heteroskedastic consistent estimator for the standard errors.

We begin the next section by discussing the basic assumptions underlying OLS. In particular, we focus on the relationship between these assumptions and the way in which complex samples are drawn. We then digress and discuss the construction of the Monte Carlo experiments presented in this article. We use the results of these experiments to illustrate the implications of using the different strategies presented here. Next, we turn to the core of the article and examine the appropriateness of using OLS or OLS with WOLS in different situations. We conclude by providing a simple set of guidelines for how and when to use sampling weights in regression.

ASSUMPTIONS BEHIND REGRESSION

The standard regression model as estimated by OLS can be developed from a number of different perspectives (e.g., see White 1984; Goldberger 1991). Because sociologists are typically interested in estimating the causal effects of a set of independent variables on a dependent variable, we assume that a regression equation represents a behavioral model, albeit possibly a very crude one, that generates the observed data. Our objective then is to estimate the causal effects of different independent variables in our regression model.[2]

Without loss of generality, we assume that there is a dependent variable Y and two independent variables X_1 and X_2. For the purposes of this article, we will assume that our data consist of a cross section of N individuals. The critical assumptions needed for OLS are

1. $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + e$ is the true model for all members of the population,
2. the X_1 and X_2 are fixed and have positive variance,
3. X_1 and X_2 are not linearly dependent,
4. X_1 and X_2 are uncorrelated with e, and
5. the e_is are independently distributed with mean 0 and variance σ_e^2.

If these assumptions hold, then the OLS estimator for β_1 and β_2 exists, is unbiased, consistent, asymptotically normally distributed, and efficient among the class of linear unbiased estimators (the OLS estimator is said to be BLUE).[3]

The assumptions that we present are sufficient to derive the OLS estimator $\hat{\beta}_{ols}$ and some of the properties of the distribution of $\hat{\beta}_{ols}$. Although additional assumptions would allow us to make stronger statements about $\hat{\beta}_{ols}$, they have little consequence for the discussion here. For example, if we make the stronger assumption that the e_is are normally distributed, then the OLS estimator will be normally distributed even for small samples, and it will be efficient with respect to the class of *all* estimators, not just linear unbiased estimators.

None of these assumptions makes any direct mention of how the data are sampled. However, different sampling schemes may well affect the validity of the last four assumptions.

The first assumption embodies the idea that a regression equation can be thought of as a structural or causal model of behavior that applies to all individuals in the population of interest. Assumption 1 will not hold if the model is missing terms that belong in the model and are correlated with included variables. This includes linear, nonlinear, and interaction terms (which are a type of nonlinear term). The question of whether the model being estimated is correctly specified is critical to understanding the potential effect of sampling weights on our regression estimates.

The assumption that the Xs are fixed means that the sample is drawn knowing beforehand the values of the X variables for individuals in the population. In general this is not true, and the Xs are in fact stochastic. This creates a set of technical issues that are of no interest or consequence here. All the results stated below are true for stochastic Xs conditional on the actual sample of Xs drawn.

The second part of assumption 2 is that the sample must be drawn so that the Xs have some variance: Individuals in the sample must be observed with different values on the X variables. If every person in the sample has the same value for a given X variable, that variable will have zero variance and it will be impossible to estimate the effect of that X on the outcome variable Y.[4]

The third assumption means that no one X can be a perfect linear combination of the other Xs in the model. If some X is a linear combination of other Xs included in the model (perfect multicollinearity), it will be impossible to distinguish between the effects of those

Xs on the outcome variable Y. In this case, the ordinary least squares estimator does not exist.

Assumptions 2 and 3 are directly related to the design of the sample. As the sample design approaches either extreme case (small variance of X or high levels of multicollinearity among the Xs), the parameter estimates yielded by the sample will remain unbiased and consistent, but they will have larger and larger standard errors. Because assumptions 2 and 3 refer to properties of the sample rather than the population, they are easily tested with sample data.

The fourth assumption, that e and X are uncorrelated, is the core assumption in regression. This assumption cannot be directly tested. If it fails to hold, the OLS estimates will be biased and inconsistent. Omitted Xs, incorrect functional specification, measurement error in the Xs, endogenous Xs, and sample selection bias can all cause this assumption to fail.

Assumption 5 states that the errors in the model are independent of each other and distributed with equal variances. In general, with multistage stratified cluster samples these conditions will not hold. Clustering often leads to a lack of independence among the errors within clusters. This has two consequences. First, OLS estimates will be unbiased and consistent, but will no longer be BLUE. Second, OLS standard errors will be wrong. Unfortunately, space constraints do not allow us to discuss methods for handling these situations here.[5]

Assumptions 2 through 5 are directly related to how the data being studied have been sampled. However, none of these assumptions requires that the distribution of the X variables in the sample be similar to that in the population. Figure 1 illustrates the true (population) relationship between a single X and Y. If, on the one hand, a sample has been drawn with X values of A and B, the slope and intercept could be estimated by drawing a line through the average value of Y at these two points. This is what OLS would do. If, on the other hand, individuals with X values of C and D had been sampled, the same analysis would, on average, yield the same parameter estimates. This example depends critically on the assumption that the relationship between Y and X is correctly specified.

Now assume that the true relationship between Y and X is curvilinear as depicted in Figure 2. If a linear relationship between Y and X is mistakenly posited, how individuals are sampled with respect to X can have a large effect on parameter estimates. If individuals are sampled at points A and B, a steep slope will be estimated. Alternatively, if individuals are sampled at points C and D, the estimated slope will be quite a bit flatter.

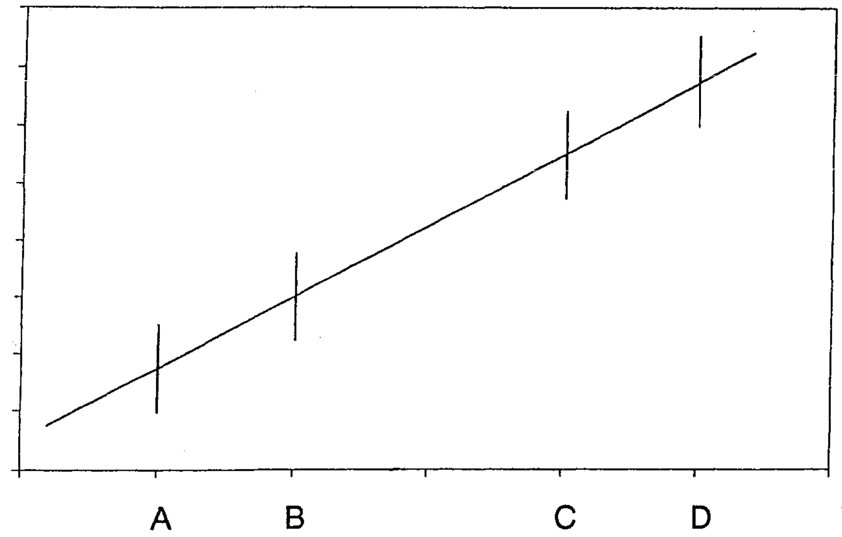

Figure 1: Effect of Different Sampling Strategies on Estimating a Linear Regression When the True Relationship Is Linear

As this example illustrates, if the model being estimated has been properly specified (whether it be linear or nonlinear), samples with different distributions of the X variables will yield (on average) the same OLS estimates. There is therefore no need for the sample distribution of the X variables to reflect the population distribution. However, if the model being estimated is misspecified, then different distributions of the Xs (whether they are produced from different samples or from a single sample with and without sampling weights) can produce OLS estimates with different expected values.

Although a correctly specified model will provide consistent and unbiased parameter estimates regardless of how a sample is drawn with respect to X, there are other factors to be considered. Central among these is that the sample yield estimates with standard errors as small as possible. In general, the smaller the variance of X in the sample, the larger the standard error of its parameter estimate. Fur-

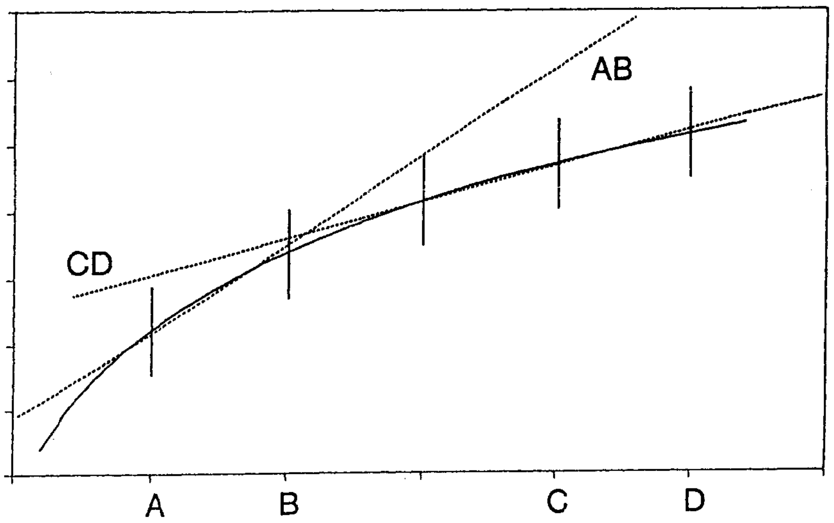

Figure 2: Effect of Different Sampling Strategies on Estimating a Linear Regression When the True Relationship Is Nonlinear

thermore, if the Xs in the model are close to being collinear, at least some of the regression coefficient estimates will have large standard errors.

Thus, although estimated regression coefficients will be unbiased and consistent regardless of how a sample is drawn with respect to the Xs, sample design can have an effect on the standard errors of those estimates. In fact, many specialized studies use sample designs that attempt to minimize standard errors of parameters of interest by oversampling groups of particular interest (Blacks, the aged, the very poor, or the very wealthy, for example). This oversampling increases the variance and may decrease the level of collinearity of some key Xs. Oversampling necessarily leads to sample distributions of some variables being quite unlike their population distributions.

MONTE CARLO EXPERIMENTS

Researchers often recognize that there may be problems with the models and methods they are using. However, many also assume (a) that the problems are of little consequence to their study, (b) that the use of more appropriate techniques would be difficult to implement, and (c) that the appropriate methods would lead to similar, if not identical, conclusions if they were used.

We believe that the methods discussed here can make a substantive difference in the conclusions that researchers might draw from a set of data. We also believe that the alternative procedures we propose are no more difficult to use than standard methods. To support these arguments, we do two things. In the following sections of this article, we first offer a mathematical presentation of the methods under consideration. From these discussions, we provide some formal bases for the procedures we recommend. We then illustrate differences between common practice and our proposed alternative by presenting the results of a series of Monte Carlo experiments.

Hidden in the language of the last section is the idea that the βs are estimated from a sample of data. That sample is only one of infinitely many different samples that could be drawn from the larger population, and each sample would yield different estimates of the βs. When we say that the OLS estimate of β is unbiased, we mean that the average estimated $\hat{\beta}_{ols}$ across infinitely many samples equals the true β. The standard error of $\hat{\beta}_{ols}$ is an estimate of the standard deviation of all of the infinitely many $\hat{\beta}_{ols}$ estimated from those samples. The properties of the estimators that we have discussed are based on mathematical properties that are derived from a set of assumptions. The question remains open whether the assumptions are reasonable and, consequently, whether the derived properties apply in any given case. Although time, space, and budget constraints prevent us from drawing infinitely many samples, we can conduct experiments that mimic that process.

Our Monte Carlo experiments involve the following steps:

1. We create a set of data. We do this by constructing a set of 2,000 individuals with a variety of X values and predicted Ys.
2. We then generate a separate random error for each of 2,000 observations in our data set and this error is added to each person's predicted Y.
3. Using these data, we estimate the model parameters. The estimation procedure will depend on the specific experiment being conducted.
4. We record those estimates.

5. We then go back to step 2 and repeat the process until we have created 750 data sets and 750 associated sets of parameter estimates.

By virtue of the fact that we have designed the experiments ourselves, we know what the true relationships in the data are before we begin estimating any models. We can therefore compare the average estimate of $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}_{est}$) to the true $\boldsymbol{\beta}$ and assess whether $\hat{\boldsymbol{\beta}}_{est}$ appears to be an unbiased estimate of $\boldsymbol{\beta}$. We can also measure the observed standard deviation of $\hat{\boldsymbol{\beta}}_{est}$ (which is a direct measure of its standard error) and compare this either to its theoretical standard error or to the standard error of some other estimate of $\boldsymbol{\beta}$. The former comparison tells us whether the theoretical standard error is an unbiased estimate of the true standard error, whereas the latter comparison tells us which estimate is more efficient (i.e., has a smaller standard error). Taken together, this information allows us to assess how often various estimation practices lead us to correct and incorrect conclusions.

One common objection to the use of evidence from Monte Carlo experiments by practicing data analysts is that because we are able to design our samples in any way we choose it is possible for us to use a sample design that makes the differences between the techniques we are criticizing and those we are advocating appear to be more extreme than they might with data from an actual survey. We are sensitive to these concerns and have gone to some lengths in order to work with data similar to the data commonly used by sociologists.

The original data for our models are taken directly from the General Social Survey (GSS) 1974-84 cumulative file. We are interested in estimating a simple wage model of the following form:

$$\text{Ln(Wage)} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\text{*Educ} + \boldsymbol{\beta}_2\text{*Black} + \boldsymbol{\beta}_3\text{*(Educ*Black)} + \boldsymbol{\beta}_4\text{*Ability} + \text{e}.$$

The GSS does not contain a variable that directly measures a respondent's wages. It does, however, have variables that assess whether or not a respondent was employed full-time and what the respondent's income was for the previous year. We worked with all men between the ages of 25 and 55 who were employed full-time in the year preceding the survey. The wage variable in our model is actual annual income for these men.[6] The GSS also includes a 10-word vocabulary test that suffices (for our purposes, at least) as a general ability

measure. Using listwise deletion of missing cases, we arrived at the following OLS estimates (standard errors are given in parentheses):

$$\begin{aligned}\text{LN(Wage)} &= \underset{(.065)}{8.529} + \underset{(.00488)}{.0461}\text{*Educ} - \underset{(.0369)}{.165}\text{*Black} \\ &+ \underset{(.0120)}{.023}\text{*(Educ*Black)} + \underset{(.00638)}{.024}\text{*Ability} + \underset{(.00135)}{.017}\text{*Age} + \underset{(.0609)}{.0609}\text{*Time}\end{aligned}$$

$$R^2 = .316 \qquad N = 1{,}601 \qquad \text{Var(e)} = .2235.$$

Using these estimates, we constructed a sample with 1,000 Black men and 1,000 White men, all age 30, earning 1980 dollars, with between 8 and 16 years of education. Within these variable ranges, we imposed the multivariate distribution of Xs found in the GSS on constructed data. We then created the following "true" model:

$$\text{Ln(Wages)} = 9.49 + .046\text{*Educ} - .165\text{*Black} + .023\text{*(Educ*Black)} + .024\text{*Ability} + \text{e}.$$

Using this equation and individuals' X values, we calculated predicted wages for all 2,000 individuals. We then created 750 independent samples of these individuals by drawing 2,000 separate errors for each sample (normally distributed with a variance of 0.25) and adding the errors to the predicted Ys. These samples are the data for the experiments we report below.

Although we would not want to use this wage model as the basis for further substantive research or policy recommendations, we believe that it is quite sufficient for our purposes. The coefficients and their standard errors are all of a similar magnitude to those that commonly appear in the regression models estimated by sociologists. Furthermore, the covariance matrix of the X variables in the models is taken directly from a commonly used data source.

THE CONSTRUCTION AND USE OF SAMPLING WEIGHTS

Before discussing the wisdom of using sampling weights in regression analysis, it is important to understand their intended purpose and how they are constructed. The purpose of sampling weights is to make the distribution of some set of variables in the data approximate the distribution of those variables in the population from which the sample was drawn.

The distribution of variables in an unweighted sample can differ from that in the population from which it was drawn for two reasons. First, individuals may be sampled with unequal probabilities. For example, a sample with a greater proportion of Blacks than are present in the population as a whole will generally be more efficient for drawing conclusions about racial differences than a representative sample.[7]

A sample may also differ from the population from which it was drawn because of random chance. For example, although a population might be 50% female, drawing a sample of individuals with equal probability could produce a sample that is only 47% female. Sampling individuals with equal probability does not insure that the resulting data will be representative of the population.

Sampling weights are typically constructed in two stages. Preliminary weights are first constructed using information about the design of the sample and response rates. These preliminary weights are approximately equal to the $1/p_i$, where p_i is a first estimate of the probability of being in the final sample. In the second stage, these preliminary weights are adjusted through a process of poststratification. This adjustment is necessary because the information used in the first stage is imperfect and, thus, there is no guarantee that the sample will be representative of the population along important dimensions even after using the first stage weights. Poststratification is a relatively simple adjustment procedure that uses population-based information (typically census data adjusted using vital statistics) to force the sample to be representative of the population along certain key dimensions (typically, age, race, and sex). The first stage weights are adjusted so that each type of individual (e.g., White women between the ages of 25 and 30) represents the same proportion of the final weighted sample as of the population.

As constructed, sampling weights are useful (often essential) for obtaining unbiased estimates of univariate population characteristics from sample data. Discussion of the use of sampling weights in this context can be found in any good sampling textbook (e.g., Kish 1965; Cochran 1963).

Many individuals' intuition would probably suggest that weighted data should also be used for the estimation of regression models. After all, using the weighted data will give covariance and variance estimates that are unbiased and consistent estimates of quantities in the population. Thus the regression estimates for the weighted sample should also be unbiased and consistent estimates of the regression model for the entire population. (We refer to this procedure as weighted ordinary least squares or WOLS in order to differentiate it from weighted least squares or WLS. Formulas for these estimators and their standard errors can be found in the appendix. In WLS the sample is also reweighted, but the weights are the inverse of each individual's error variance, $W_i = 1/\sigma_i^2$. This is done when there is heteroskedasticity in order to get efficient estimates.)

The above logic as to why one should use WOLS has appeal, but fails to take account of a number of issues. First, if the errors in the unweighted sample were homoskedastic (with variance σ_e^2, consistent with Assumption 5 above), using sampling weights will induce heteroskedasticity. This is because weighting is equivalent to multiplying each observation by W_i. If an individual's error term was originally e_i, after weighting it will be $W_i e_i$ and the variance of their error term will be $\sigma_{ei}^2 = W_i^2 \sigma_e^2$.

A different type of situation arises when researchers, in the interest of compressing large data sets, create weights to indicate the number of individuals with each observed combination of values on the variables of interest. For instance, if there are three Black women, all age 25, with high school diplomas and annual earnings of $20,000 in a data set, only one such case will be stored in the data set, but it will be given a weight of 3. (Some packages refer to these as frequency weights.) In this case, most statistical packages will calculate the standard errors correctly. This is not what is being done with sampling weights. With sampling weights there is a single individual in the sample. The statistical package calculates the wrong standard errors because it assumes that there are W_i individuals in the sample. Equivalently statistical packages produce the wrong standard errors because the formulas they use assume that in the weighted sample all individuals have the same error variances where in fact (assuming there is no heteroskedasticity) the error variances are $W_i^2 \sigma_e^2$. This is perhaps the greatest danger in using WOLS.

We show in our Monte Carlo experiments that the standard errors reported by statistics packages when WOLS is used can be biased in either direction, resulting in erroneous hypothesis tests (both false positives and false negatives) and confidence intervals of incorrect width.[8] If sampling weights are to be used, a formula for the coefficient standard errors that takes account of the heteroskedasticity is needed. We will show how White's (1980) heteroskedastic consistent estimator can be used to do this.

Assuming that standard errors for WOLS can be correctly calculated, there is the further question of whether to use OLS or WOLS. There are two cases to consider. The first is where the weights, W_i, are a function of Xs that are included in the regression model. The other situation is where the weights are a function of not only the Xs in the model but also the dependent variable.

WHEN SAMPLING WEIGHTS ARE A FUNCTION OF INDEPENDENT VARIABLES

When sampling weights are only a function of independent variables included in the model being estimated, unweighted OLS will be the appropriate course to take. In this case, using or not using weights is analogous to drawing samples with different distributions of the independent variables. We have already illustrated how when the model being estimated is correctly specified parameter estimates will be unbiased and consistent regardless of the distribution of the independent variables in the sample (as long as the variables have positive variance and are linearly independent). This means that both OLS and WOLS will yield unbiased and consistent estimates. The Gauss-Markov theorem, however, guarantees that OLS will be more efficient, yielding smaller standard errors. As a result, OLS is to be preferred over WOLS.

Table 1 presents the results of using OLS and WOLS to estimate our model predicting log wages. Because our sample was constructed to be composed of 50% White and 50% Black respondents, sampling weights were calculated to produce a weighted sample composition of 10% Black and 90% White respondents (weight for Whites = 1.8, weight for Blacks = 0.2), reflecting their approximate proportions in

TABLE 1: Monte Carlo Results for Properly Specified Model

	Intercept	*Education*	*Race*	*Race* Education*	*Ability*
Expected value	9.4900	0.0460	–0.1650	0.0230	0.0240
Theoretical OLS standard error	0.0501	0.0081	0.0232	0.0112	0.0083
Average OLS estimate	9.4920	0.0458	–0.1660	0.0229	0.0235
Observed standard error	0.0498	0.0079	0.0237	0.0108	0.0084
Average package standard error	0.0498	0.0080	0.0230	0.0111	0.0083
Average WOLS estimate	9.4900	0.0458	–0.1660	0.0229	0.0237
Observed standard error	0.0650	0.0081	0.0240	0.0108	0.0112
Average correct standard error	0.0639	0.0083	0.0234	0.0111	0.0109
Average package standard error	0.0512	0.0063	0.0385	0.0192	0.0087

NOTE: 750 samples were generated using the following equation:

Ln(Wages) = 9.49 + (0.046 * Educ) –
(0.165 * Black) + (0.023 * Educ * Black) + (0.024 * Ability) + e.

e is a normal random variate with a variance of 0.25.

Expected values of the parameters are the coefficients. Theoretical standard errors are based on the standard OLS formula ($\sigma^2(X'X)^{-1}$). Average estimate is the mean parameter estimate across 750 generated samples. Observed standard error is the standard deviation of the parameter estimate across 750 generated samples. Average correct standard error is the mean standard error, using the theoretically correct formula discussed in the text, across 750 generated samples. Average package standard error is the mean standard error, using the default package formula, across 750 generated samples. Education is recoded so that high school (12 years) is centered at 0. Race is coded 1 for Blacks, 0 for all others. Ability is a 10-point scale from the General Social Survey 10-item vocabulary test (WORDSUM).

the national population of the United States. The expected values for the correctly specified model (the true model) are presented at the top of the table.

The first thing to notice is that when the correctly specified model is estimated, OLS parameter estimates match their expected values. Additionally, the observed standard errors of those parameters are quite close to the theoretical standard errors.

When weights are applied to the data, the mean value of the regression parameters is also close to their true values. There are, however, three important changes. First, the standard errors produced by the regression program are wrong on average. This can be seen by comparing the average package standard errors to the observed standard errors of the parameter estimates. If we could predict the direction and the approximate magnitude of these mistakes, things might not seem so bad. However, our results show that the bias can be either positive or negative. For example, the education coefficient's standard error is too small (.00634 instead of .00827), but the race coefficient's standard error is too large (.0192 instead of .0111). In general, it will not be possible to predict the direction of this bias. Furthermore, in this case the biased standard errors could lead an analyst to incorrectly exclude the race-education interaction from the model.

When the correct standard errors for the weighted regression model are used, it is clear from Table 1 that use of sampling weights will generally yield less efficient parameter estimates than use of the unweighted sample. For example, the true OLS standard error for Ability is .00833, whereas the true WOLS standard error is .0109 (an increase of 31%). A loss in efficiency of this magnitude could easily lead to the incorrect rejection of variables from a model being estimated, even using the correct standard errors. About the only thing to be said for the use of sampling weights in this correctly specified model is that the slope estimates themselves are unbiased.

If the parameter estimates from the weighted and unweighted samples differ (that is, if the parameter estimates are sensitive to how the sample is drawn with respect to X), this is an indication that the model being estimated is not correctly specified or that the weights are a function of the dependent variable. We discuss this second possibility later in the article.

There are several ways to understand why sensitivity of parameter estimates to the use of sampling weights can be an indication of model misspecification. One way to view the situation is as a problem of pooling two separate samples: one sample of White respondents and a separate sample of Black respondents. The outcomes for each group (wages) are the result of different underlying processes that can be described by the following two models:

White Model

$$\text{Ln(Wages)} = 9.490 + .046^*\text{Educ} + .024^*\text{Ability} + e.$$

Black Model

$$\text{Ln(Wages)} = 9.325 + .069^*\text{Educ} + .024^*\text{Ability} + e.$$

We are interested in combining these two samples in order to estimate a single model. The correctly specified model would be

$$\text{Ln(Wages)} = 9.49 + .046^*\text{Educ} - .165^*\text{Black} + .023^*(\text{Educ}^*\text{Black}) + .024^*\text{Ability} + e.$$

Because the Educ*Black interaction is correlated with the other X variables in the model, omitting it from the model would yield estimates for the remaining parameters that would be biased and inconsistent (see Hanushek and Jackson 1977, pp. 79-86). Leamer (1978, p. 76) shows that the estimates obtained using the pooled sample will be a weighted average of the estimates for the two separate samples. The weight referred to by Leamer is a function of the sample covariance matrices of the Xs for each group and the size of each sample. By implying different relative sizes for the two groups, regression estimates that do and do not use sampling weights may give very different results.

This analysis can also be applied when a nonlinear relationship has been misspecified. When this is the case (as in Figure 2), there are some individuals for whom the effect of the independent variable is large and others for whom the effect is smaller. The size of the OLS estimate will depend critically on how many individuals of each type are in the sample.

Sensitivity of parameter estimates to the use of sampling weights can also be an indication that a variable has been omitted from the model. Suppose that the correctly specified model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e.$$

Instead, a model is estimated that omits X_2:

$$Y = B_0 + B_1 X_1 + e.$$

The expected value of the coefficient for X_1 in this misspecified model, B_1, will be

$$E(B_1) = \beta_1 + s_{21}\beta_2,$$

where s_{21} is the regression slope for X_2 regressed on X_1. Note that s_{21} is solely a function of the Xs as it is equal to

$$s_{21} = \frac{\text{cov}(X_2, X_1)}{\text{var}(X_1)}.$$

Thus the size of the bias in B_1 will depend on s_{21}. This, in turn, depends on the variance of X_1 and the covariance of X_2 and X_1. These last two quantities, and therefore the value of s_{21}, will generally be sensitive to sampling weights. This can be seen from the formulas for the weighted and unweighted quantities:

$$\frac{\Sigma(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{n} = \text{Cov}(X_1X_2)$$

$$\frac{\Sigma W_i(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\Sigma W_i} = \text{Weighted Cov}(X_1X_2)$$

$$\frac{\Sigma(X_{i1} - \bar{X}_1)^2}{n} = \text{Var}(X_1)$$

$$\frac{\Sigma W_i(X_{i1} - \bar{X}_1)^2}{\Sigma W_i} = \text{Weighted Var}(X_1).$$

The weighted values will generally only equal the unweighted values when all persons have the same weight. This means that typically the only time that the WOLS estimate of B_1 equals the (unweighted) OLS estimate of $\boldsymbol{\beta}_1$ is when s_{21} is zero, that is, when there is no omitted variable bias. All other times, the size of s_{21} (the size of the bias) will be influenced by the presence or absence of sampling weights.

Thus evidence that the WOLS estimate of $\boldsymbol{\beta}_1$ and the (unweighted) OLS estimate of $\boldsymbol{\beta}_1$ are different suggests that the model being estimated suffers from omitted variable bias. The omitted variables may be nonlinear (including interaction) terms or other Xs that need to be included in the model.[9]

Table 2 summarizes the results of our second round of Monte Carlo experiments illustrating the effects of sampling weights in a misspecified regression model. When our incorrectly specified model (we have omitted the Race-Education interaction effect) is estimated with OLS, the results are as we would expect: Because the variables included in our model are correlated with the omitted variable, their estimated coefficients are biased. The estimated education effect is about midway between the true White and Black education effects (reflecting the fact that Blacks and Whites make up equal portions of our sample). Because the correlation between the omitted interaction variable and the Ability measure is relatively small, the bias in the estimated Ability coefficient is not large.

TABLE 2: Monte Carlo Results for Misspecified Model[a]

	Intercept	Education	Race	Race * Education	Ability
Correctly specified model					
Expected value	9.4900	0.0460	−0.1650	0.0230	0.0240
Theoretical OLS standard error	0.0501	0.0081	0.0232	0.0112	0.0083
Misspecified model					
Average OLS estimate	9.4880	0.0569	−0.1580		0.0232
Observed standard error	0.0499	0.0061	0.0234		0.0084
Average package standard error	0.0503	0.0060	0.0230		0.0084
Average WOLS estimate	9.4900	0.0479	−0.1600		0.0236
Observed standard error	0.0650	0.0075	0.0237		0.0112
Average correct standard error	0.0652	0.0078	0.0236		0.0111
Average package standard error	0.0513	0.0061	0.0383		0.0088

NOTE: Averages are across 750 samples. See notes to Table 1 for definitions.
a. Missing Race × Education interaction.

As in the case with OLS, our misspecified model has produced a biased estimate of the education effect. With WOLS, however, the bias has a somewhat more appealing quality: If we were interested in describing the average monetary return to education for people in the United States, this estimate would reflect the average payoff in a population that is 90% White and 10% Black. However, if the interest is in a structural (causal) interpretation of the parameter estimates, this coefficient would have little meaning.[10]

Estimating the incorrectly specified regression model with sampling weights shows that the regression package produces incorrect standard errors. As before, some of these errors are too small and some are too large. Furthermore, the estimated parameters have larger (correct) standard errors than the OLS estimates for the misspecified model, again demonstrating that WOLS is less efficient.

MODEL RESPECIFICATION

The above analyses suggest that if OLS and WOLS yield similar results, the OLS estimates are preferable because they have smaller standard errors. DuMouchel and Duncan (1983) suggest a simple procedure to determine whether estimates from the weighted and unweighted models are significantly different and why. They propose estimating the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \delta_0 W + \delta_1 (W^*X_1) + \delta_2 (W^*X_2) + e,$$

where X_1 and X_2 are independent variables and W is the weight variable. (The model of interest can have any number of independent variables.)

If the original model is properly specified, then the weight variable and the interaction of it with the X variables should explain no additional variance in Y, and $\delta_0 = \delta_1 = \delta_2 = 0$. A standard *F* test is carried out for the difference in explanatory power between the two models. The procedure advocated by DuMouchel and Duncan (1983) thus consists of testing for the significance of a variable W that is presumed to be a function of possibly omitted Xs and the interaction of W with the included Xs. If the *F* test is not significant, then the weighted and unweighted estimates are not significantly different and the analyst can proceed by using unweighted OLS. Weighted and unweighted estimates are significantly different if the *F* test is significant.

If the OLS and WOLS estimates are indeed different, the specific coefficients from the augmented model can be examined relative to their standard errors for possible sources of misspecification. There are two outcomes of interest. First, coefficients involving the terms with the weight variables may be large and significant, suggesting the possible need for additional variables and/or interactions and nonlinear terms involving these Xs. Second, a coefficient may be small, but significant, typically as the result of a very large sample. Here respecification may only be appropriate if additional variables, nonlinear terms, or interactions added to improve the model's specification are of substantive interest.[11]

In the example presented in Table 2 (where the Race*Education interaction has been dropped), the DuMouchel and Duncan (1983) procedure would involve estimating the following model:

$$\text{log wage} = \beta_0 + \beta_1\text{race} + \beta_2\text{education} + \beta_3\text{ability} + \beta_4\text{weight} + \beta_5(\text{weight*race}) + \beta_6(\text{weight*education}) + \beta_7(\text{weight*ability}).$$

There are two reasons that the WX terms might have a large or significant effect. First, the weight variable itself might have an effect (here measured by β_4). This would happen if the weights were a function of Xs that had been omitted from the equation. Second, there may be important interaction terms missing from the model involving the X variables for which the interactions between the weight variable and the other Xs are a proxy.

In our proposed analysis, error messages from our statistical software would quickly alert us to the fact that it is not possible to estimate the coefficient for the weight variable (β_4) and for the Weight*Race interaction because they are both perfectly collinear with race. Omitting these two terms and fitting a model with the remaining variables would produce a significant coefficient for the interaction of the weight variable with education. This would suggest the possible need to include some type of interaction involving education. Because sampling weights are most often functions of age, race, and sex, we might want to experiment by including age and race (although not sex, because this sample contains only men), and their interactions with education. We might also want to try an interaction between education and ability if we thought the weight variable was related to ability, although in most surveys this would be unlikely. A third possibility would be to consider omitted variables that the Weight-Education variable might be acting as proxy for. In our model in the end, we would decide to respecify the model by adding an interaction between race and education.

WHEN SAMPLING WEIGHTS ARE A FUNCTION OF THE DEPENDENT VARIABLE

There will be situations where OLS and WOLS parameter estimates differ, where the DuMouchel and Duncan *F* test is significant, but attempts to respecify the model by including additional independent variables (short of the sampling weights themselves) fail to solve the problem. These are likely to be instances where the probability of an individual being sampled, and thus the sampling weights, are a function of the dependent variable. These cases generally arise directly from the design of the sample being used. To cite just three specific examples, the University of Michigan's Panel Study of Income Dynamics (PSID) has an over-sample of people who lived in poor families in the mid-1960s, the Federal Reserve Board's Survey of Consumer Finances has an over-sample of families with very high incomes, and current plans for the 1995 redesign of the Census Bureau's Survey of Income and Program Participation (SIPP) involve over-sampling households located in areas that had high poverty concentrations at the time of the 1990 decennial census. In each of these examples, the sampling weights are (or will be) a function of income, a common dependent variable.[12]

To illustrate how sampling weights can affect parameter estimates when they are a function of the dependent variable, consider the problem of estimating the effect of education on income where the only individuals in the sample are those with incomes below $15,000. (Hausman and Wise 1977 provide a similar example.) Figure 3 illustrates the situation.

If the model is given by

$$\text{Income} = \beta_0 + \beta_1 \text{ Education} + e,$$

then individuals can have incomes over $15,000 because they have high education or because they have a large e. By restricting the sample to those with observed incomes under $15,000, individuals with large positive error terms have been eliminated from the sample. This has, in turn, induced a negative correlation between education and the error term (e): For those who remain in the sample, the more education a person has, the less likely it is that they have a positive e and, equivalently, the more likely it is that they have a (large) negative e. This is an example of a truncated sample that is an extreme version of sample selection bias (Berk 1983; Winship and Mare 1992), and it results in parameter estimates that are biased (generally toward zero) and inconsistent. (For a general discussion of the relation between truncation, censoring, and sample selection bias, see Greene 1990, chap. 21.)

If instead of observing no individuals with incomes over $15,000, individuals with incomes over $15,000 were half as likely to be sampled as individuals with incomes below that amount, education

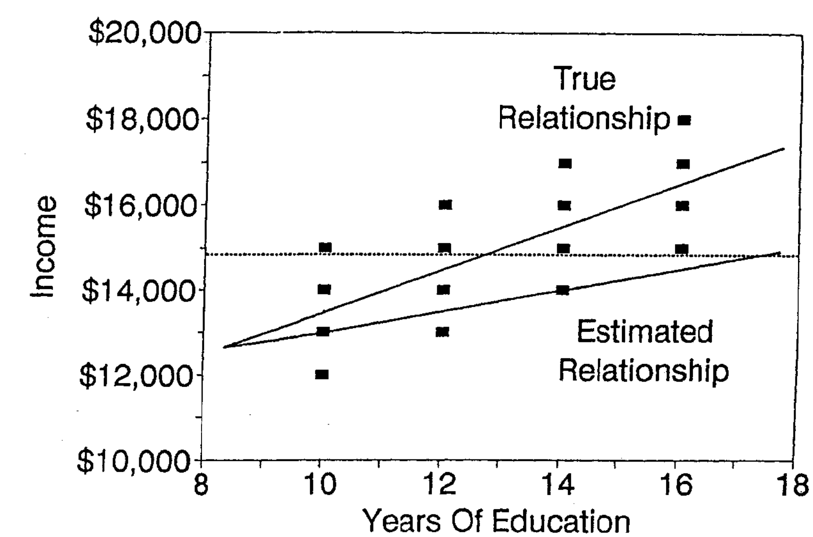

Figure 3: Estimating the Effect of Education on Income From a Sample of Persons With Incomes Below $15,000

would still be negatively correlated with the error term, although not as severely. In this case, however, use of weights that reflect the differential sampling based on income would help to reduce that negative correlation between education and the error term. More generally, in the situation where sampling weights are a function of Y, WOLS provides consistent estimates of the true regression slopes.[13]

As in the general case, the use of WOLS will induce heteroskedasticity in the error terms. The structure of the variances of the error terms will typically depend on the nature of the sampling scheme. Rather than trying to explicitly model the variance structure, a simpler approach is to use the White heteroskedastic consistent estimator for the standard errors (White 1980). White shows that, although generally one cannot consistently estimate the variance of each individual's error, it is possible to estimate the true standard errors using the estimated residuals. Several programs including SAS, STATA, and LIMDEP have routines that provide White heteroskedastic consistent standard errors.[14]

We demonstrate this use of weights in Table 3, where we have assumed that the probability that individuals are selected into the

TABLE 3: Monte Carlo Results Illustrating the Use of Sampling Weights in the Presence of Sample Truncation[a]

	Intercept	*Education*	*Race*	*Race* Education*	*Ability*
Expected value	9.4900	0.0460	–0.1650	0.0230	0.0240
Theoretical OLS standard error	0.0501	0.0081	0.0232	0.0112	0.0083
Full sample					
Average OLS estimate	9.4934	0.0460	–0.1663	0.0230	0.0234
Observed standard error	0.0505	0.0076	0.0230	0.0110	0.0083
Average package standard error	0.0498	0.0080	0.0231	0.0111	0.0083
Truncated sample					
Average OLS estimate	9.3059	0.0385	–0.1303	0.0151	0.0188
Observed standard error	0.0523	0.0092	0.0244	0.0126	0.0086
Average package standard error	0.0537	0.0090	0.0249	0.0122	0.0090
Average WOLS estimate	9.4933	0.0457	–0.1659	0.0236	0.0232
Observed standard error	0.0736	0.0118	0.0334	0.0169	0.0119
Average package standard error	0.0603	0.0097	0.0280	0.0135	0.0100
Average White standard error	0.0727	0.0117	0.0337	0.0161	0.0121

NOTE: Averages are across 750 samples. See notes to Table 1 for definitions.
a. Of cases with annual wages over $15,000, 70% are missing from the sample.

sample is a function of their income. Specifically, we assume that persons with annual wages over $15,000 are only 30% as likely to be sampled as those with annual wages under $15,000. Such a situation could arise in practice in a survey that attempted to oversample those with low wages or in a survey with low participation rates among those with higher wages.

The results clearly show that attempting to estimate the model without correcting for sample selection yields biased estimates of the βs. The average OLS estimates from the truncated sample are all biased toward zero, with the average estimated Race*Education coefficient being only 66% as large as the true effect. Using WOLS drastically improves the coefficient estimates: The average WOLS estimates of β are all within 2% of their true values. However, the average estimated standard errors produced by the statistical software are all only about 80% of the observed standard errors of the coefficients. In general these miscalculated standard errors could also be larger than their true values. What should be clear from this example is that the standard errors generated for weighted data by most statistics packages can be off by enough to affect formal hypothesis tests as well as decisions about model specification. Table 3 also provides the White heteroskedastic consistent standard errors. These are quite close to the observed standard errors.

CONCLUSION

When a researcher is going to perform a regression analysis with data that have sampling weights, what should be done? First, the analyst should estimate two models: one with unweighted data (OLS) and one using the sampling weights (WOLS). If the parameter estimates are substantively similar, then the OLS estimates are preferable because they are more efficient and the estimated standard errors will be correct. If in doubt about whether the OLS and WOLS estimates are different, the *F* test proposed by DuMouchel and Duncan (1983) can be easily performed. As we stated before, if the data come from a clustered sample and nothing is done to correct for this, caution should be used in interpreting the *F* test. In addition, caution is needed if the researcher carries out multiple significance tests.

When OLS and WOLS produce different parameter estimates, the researcher needs to carefully consider the possible reasons. One possibility is that the model may be missing linear, nonlinear, or interaction terms. Estimating DuMouchel and Duncan's augmented equation and examining the effects of the weight variable or interactions of it with other variables provides a way of diagnosing misspecification.

If respecifying the model does not make weighted and unweighted estimates similar, the other possibility is that the weights are correcting for sample selection bias. In this case, use of the weights (WOLS) will yield consistent parameter estimates, but incorrect standard errors. Consistent estimates of the standard errors can be gotten using White's estimator.

APPENDIX
The OLS Estimator in Matrix Notation

Given that Y and e are (n × 1) column vectors, X is an (n × k) matrix, and B and $\hat{\beta}_{ols}$ are (k × 1) column vectors. Assume

1. $Y = XB + e$ is the true model for all members of the population;
2. X is fixed;
3. $(X'X)^{-1}$ exists;
4. $X'e = 0$; and
5. $E(ee') = \sigma^2 I$, and $E(e_i) = 0$.

Then the OLS estimators for B and for Cov(B) are given by

$$\hat{\beta}_{ols} = (X'X)^{-1}(X'Y)$$

and

$$\mathrm{Cov}(\hat{\beta}_{ols}) = \sigma_e^2(X'X)^{-1}.$$

STATISTICAL PACKAGES AND THE FORMULAS THEY USE

If W is an (n × n) diagonal matrix of sampling weights, then the general formulas used by SPSS, SAS, Systat, STATA, and other standard packages for the Weighted OLS (WOLS) estimator, $\hat{\beta}_{wols}$ and $\mathrm{Cov}(\hat{\beta}_{wols})$ are the standard Weighted Least Squares (WLS) formulas given by

$$\hat{\beta}_{wls} = (X'WX)^{-1}(X'WY)$$

and

$$\mathrm{Cov}(\hat{\beta}_{wls}) = \sigma_e^2(X'WX)^{-1}.$$

There are small differences in the ways standard statistical packages modify these formulas for $\hat{\beta}_{wls}$ and $\mathrm{Cov}(\hat{\beta}_{wls})$ when weights are used. Systat, for example, uses only the integer portion of weight value. If an individual has a weight of 3.75, SYSTAT truncates the value and uses a value of 3. Although the correct number of observations for computation of standard errors and hypothesis testing is n, SPSS and SYSTAT use the sum of the weights. Where weights sum to the total U.S. population (roughly 250 million), this can clearly lead to problematic results from this error alone. STATA and SAS correctly compute the sample size when weights are used with their simplest regression routines.

If the unweighted data are homoskedastic (the first part of assumption 5 above), use of sampling weights actually creates heteroskedasticity. The correct formula for $\mathrm{Cov}(\hat{\beta}_{wols})$ is therefore

$$\mathrm{Cov}(\hat{\beta}_{wols}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

where, in this case, $\Omega = \sigma_e^2 W^{-1}$. This is clearly different from the formula above for $\mathrm{Cov}(\hat{\beta}_{wls})$.

NOTES

1. Although our discussion focuses entirely on least-squares models, all of our conclusions also apply to probit, logit, and other types of generalized linear models. See Manski and Lerman (1977) or Amemiya (1985) for a discussion of the estimation of conditional logit models from choice-based samples.

2. An alternative approach would be to interpret the regression model as describing differences in the conditional mean of Y across values of the specific set of Xs included in the model (Goldberger 1991). This kind of regression model is sometimes referred to as a descriptive model (and more formally known as the conditional expectation function). In this case, we would analyze the problem of estimating a regression model with and without weights from a population-based perspective. We would reach similar conclusions, but the motivation for the analysis would be less clear.

3. Asymptotic properties apply to large samples. How large the sample needs to be depends on the specific context. Most of the sample surveys sociologists work with are sufficiently large for this property to apply.

4. This would be equivalent to running an experiment that contained either no control group or no treatment group: In either case it would be impossible to determine the effect of the treatment on the outcome.

5. A number of different approaches can be used to handle this problem, including random effect (GLS), fixed effect estimators (Maddala 1977), jackknife methods (Efron and Tibshirani 1993), and linearized standard error formulas (Holt, Smith, and Winter 1980).

6. We assumed that all of the income received by men in this age group who were employed full-time year-round was from earnings. We could have used an hourly wage variable. This would have amounted to dividing annual income for these men by 2000. This would shift our intercept down by 7.6 (which is ln[2000]), but all other coefficients would be unchanged.

7. Nonresponse in cross-sectional surveys, and its longitudinal analogue, attrition bias, also contribute to samples diverging from their populations along key dimensions. Because survey organizations often know something (however little) about nonrespondents of both types, both are considered to be akin to sample design when weights are estimated.

8. Below, when we claim that OLS standard errors will be less than or equal to WOLS standard errors, we are referring to the *correctly computed* standard errors, not those that are automatically produced by standard software routines.

9. Kott (1991) argues that in each of these cases sample weights should be used so that the parameter estimates are at least consistent estimates of the regression function for the population.

Where the regression model parameters are being used for purposes of description, this can be a useful tack. However, there seems little advantage to this approach when the interest is in estimating the *causal* effects of a set of independent variables on a dependent variable.

10. If the goal is to knowingly obtain consistent estimates for a descriptive model (which may not be a correctly specified causal model) as applied to the population, then there is a rationale for weighting the data (see note 9). In general, however, the weights that would give efficient estimates of the parameters in the misspecified model will differ from standard sampling weights. For further discussion of the estimation of misspecified models, see White (1982).

11. Both the *F* test and formulas for the standard errors used here assume that error terms are uncorrelated across individuals. Because most commonly used survey data use cluster samples, there is potential correlation of errors across individuals in the same cluster. If this is not explicitly dealt with (by using GLS, a difference estimator, bootstrap methods, or linearized standard error formulas), the *F* test and standard errors should be interpreted cautiously.

In addition, if the researcher carries out a sequence of tests, say by examining the *t* values of a number of different variables, he or she needs to be additionally cautious in interpreting levels of significance. Leamer (1978) provides an extensive discussion of the problem of using significance tests in specification searches.

12. When income is an independent variable predicting other outcomes, models using data from these sources may well be amenable to the techniques described earlier.

13. In this case, WOLS can be interpreted as an instrumental variable estimator where WX is used as an instrument for X. Consistency of WOLS follows from the fact that WX is uncorrelated with the error term. If the errors are homoskedastic, the instrumental variable formula for the standard errors for the coefficients can be used. Typically, however, selection induces heteroskedasticity in the errors. The formula for the true standard errors will then depend on the distribution of the error term. It is useful to note that White's (1980) formula does not depend on the distribution of the error term.

14. The correct formula for the covariance matrix for the estimated coefficients when there is heteroskedasticity is $cov(\hat{\beta}) = (X'X)^{-1} X'\Omega X (X'X)^{-1}$. In the past, it was thought that it was not possible to consistently estimate this unless one could specify the structure of Ω and consistently estimate Ω. White (1980) showed that it is possible to consistently estimate $X'\hat{\Omega}X$ without knowing the structure of Ω. As a result, one can consistently estimate $cov(\hat{\beta})$ under general forms of heteroskedasticity.

REFERENCES

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Berk, Richard A. 1983. "An Introduction to Sample Selection Bias in Sociological Data." *American Sociological Review* 48:386-98.

Cochran, William G. 1963. *Sampling Techniques*. New York: Wiley.

DuMouchel, William H. and Greg J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analysis." *Journal of the American Statistical Association* 78(383):535-43.

Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Greene, William H. 1990. *Econometric Analysis*. New York: Macmillan.

Hanushek, Eric A. and John E. Jackson. 1977. *Statistical Methods for Social Scientists*. New York: Academic.

Hausman, Jerry and David Wise. 1977. "Social Experimentation, Truncated Distributions, and Efficient Estimation." *Econometrica* 45:919-38.

Holt, D., T.M.F. Smith, and P. D. Winter. 1980. "Regression Analysis of the Data From Complex Surveys." *Journal of the Royal Statistical Society (Ser A)* 143:474-87.

Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.

Kott, Phillip S. 1991. "A Model-Based Look at Linear Regression with Survey Data." *American Statistician* 45(2):107-12.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Maddala, G. S. 1977. *Econometrics*. New York: McGraw-Hill.

Manski, C. and S. Lerman. 1977. "The Estimation of Choice Probabilities From Choice Based Samples." *Econometrica* 45:1977-88.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48:817-38.

———. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1-25.

———. 1984. *Asymptotic Theory for Econometricians*. Orlando, FL: Academic Press.

Winship, Christopher and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327-50.

Christopher Winship is a professor of sociology at Harvard University. He is currently doing research on the application of latent class models to the structural analysis of discrete data, models of simultaneity for the relationship between the employment and marital status of young men, the rhetoric of poverty and race, and on the relationship between group structures and time scheduling problems. Recent publications include "Loglinear Models for Reciprocal and Other Simultaneous Effects," Sociological Methodology *(1991) and "Models for Sample Selection Bias,"* Annual Review of Sociology *(1992). Both articles are coauthored with Robert D. Mare.*

Larry Radbill is a professional staff member with the Joint Economic Committee of Congress. His primary research interests include trends in poverty, social well-being, and social welfare programs, along with related measurement issues. He is the author (with Kathleen Short) of Extended Measures of Well-Being: Selected Data from the 1984 Survey of Income and Program Participation, *U.S. Bureau of the Census, Current Population Reports, Series P-70, No. 26, 1992.*