



Taylor & Francis
Taylor & Francis Group



Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples

Author(s): William H. DuMouchel and Greg J. Duncan

Source: *Journal of the American Statistical Association*, Sep., 1983, Vol. 78, No. 383
(Sep., 1983), pp. 535-543

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: <https://www.jstor.org/stable/2288115>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

Taylor & Francis, Ltd. and American Statistical Association are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples

WILLIAM H. DuMOUCHEL and GREG J. DUNCAN*

The rationale for the use of sample survey weights in a least squares regression analysis is examined with respect to four increasingly general specifications of the population regression model. The appropriateness of the weighted regression estimate depends on which model is chosen. A proposal is made to use the difference between the weighted and unweighted estimates as an aid in choosing the appropriate model and hence the appropriate estimator. When applied to an analysis of the familial and environmental determinants of the educational level attained by a sample of young adults, the methods lead to a revision of the initial additive model in which interaction terms between county unemployment and race, as well as between sex and mother's education, are included.

KEY WORDS: Weighted regression; Finite population; Super population model; Educational attainment model.

1. INTRODUCTION

Suppose that a sample survey measures $(p + 1)$ variables on each of n individuals, so that the data consist of the $n \times 1$ matrix Y and the $n \times p$ matrix X . Then the least squares estimator of the regression coefficients of Y on X is

$$\hat{\beta} = (X'X)^{-1}X'Y. \quad (1.1)$$

However, the rows of Y and X often are not a simple random sample from the population. Differential sampling rates and differential response rates among various strata lead to different probabilities of selection for each individual. Kish (1965) discusses the computation of these probabilities for various sampling schemes, but this article is concerned only with stratified sampling and not with the further complication of cluster sampling. Further, the stratification is permitted to be based on X but not on Y .

As described in Kish (1965), the differential sampling and response rates lead to the computation of weights for

each case, which attempts to give each stratum the same relative importance in the sample that it has in the population. This article assumes that an observable stratification variable \tilde{J} takes on k levels and that $\{\pi_j\}$, the proportions of the population for which $\tilde{J} = j, j = 1, \dots, k$ are known. Let n_j be the size of a simple random sample drawn from the j th stratum, $j = 1, \dots, k$, so that $n_1 + \dots + n_k = n$. Since the j th stratum is underrepresented in the sample by a factor proportional to n_j/π_j , the weight assigned to the i th observation is

$$w_i \propto \pi_{j_i}/n_{j_i}, \quad (1.2)$$

where j_i is the value of \tilde{J} for the i th observation, $i = 1, \dots, n$. Let W denote the diagonal matrix whose i th diagonal element is w_i . In some textbooks, and in many analyses of survey data (see Klein 1953, Bachman, Green, and Wirtanen 1974, Blumenthal et al. (1972), Duncan and Morgan (1976), Hu and Stromsdorfer (1970), and Juster (1976)), a weighted least squares estimator is used, namely,

$$\hat{\beta}_w = (X'WX)^{-1}X'WY. \quad (1.3)$$

Which estimator should be used? Controversy has raged at least since Klein and Morgan (1951). The advocates of $\hat{\beta}$ can point out that the justification for weighted regression in terms of adjusting for unequal error variances (see, e.g., Draper and Smith 1966) is not at issue here. In the usual homoscedastic regression model, $\hat{\beta}$ is minimum variance unbiased whether or not the strata are sampled proportional to size. Nevertheless, the advocates of $\hat{\beta}_w$ are concerned with reducing the supposed bias caused by the sampling scheme, reasoning by analogy to the estimation of an overall population total or mean. In that case, such weighting is clearly necessary if there are systematic differences in the stratum means. In addition, they argue that the assumptions that lead to the optimality of $\hat{\beta}$ are likely to be violated in populations of interest. Brewer and Mellor (1973) discuss how the choice between $\hat{\beta}$ and $\hat{\beta}_w$ is influenced by the choice of a model-based approach to inference versus an approach based on randomization within a finite population in which no particular model is assumed.

The point of this article is to clarify this issue by showing how the appropriate estimator depends on which of

* William H. DuMouchel is Associate Professor of Applied Mathematics, Statistics Center, Massachusetts Institute of Technology, Cambridge, MA 02139. Greg J. Duncan is Senior Study Director, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48109. The authors want to thank Professor Herman Chernoff for his careful reading and constructive comments on an earlier draft of this article. They also want to thank two editors, three associate editors, and four referees for their many helpful comments on earlier drafts, and Martin Frankel, Jr., Graham Kalton, and Roger Wright for valuable discussion.

several possible regression models (if any) is appropriate and to show how a test based on $\hat{\beta}_w - \hat{\beta}$ may be used as a device to help decide which model, and hence which estimator, is appropriate. In Section 2 we define four different regression models of increasing generality that might be used to justify the use of $\hat{\beta}_w$. In section 3 we discuss the relationship between the models and the choice of estimator, and in Section 4 we show how an easily computed test based on $\hat{\beta}_w - \hat{\beta}$ may help in choosing a model. Section 5 contains further discussion, and in the last two sections we illustrate the issues by the construction of an educational attainment model based on a national survey.

2. FOUR REGRESSION MODELS

2.1 Notation

The decision of whether to use the weights depends on what one assumes about the population from which the data have been drawn. In this section we describe four models that exemplify the most common assumptions. Associated with each model is a certain target quantity, or parameter of interest. The question is whether $\hat{\beta}$ or $\hat{\beta}_w$ is the more appropriate estimate of that target quantity.

The reader may find it easier to think in terms of sampling from an infinite population, since population size per se is not a major issue here. We always assume that the stratum sample sizes $\{n_j\}$ are small fractions of the corresponding population stratum sizes, and the mathematics of sampling with replacement or from infinite populations are used throughout this article. Let \tilde{Y} and \tilde{X} denote the scalar and $(1 \times p)$ random variables defined by a single draw of the dependent and independent variables, respectively, from the entire population. Let y and x denote values of \tilde{Y} and \tilde{X} , namely, single rows of the data matrices Y and X respectively. Unconditional expectations $E(\cdot)$ refer to a simple random sample from the population, while conditional expectations $E(\cdot | J)$ refer to stratified sampling, where a simple random sample of size n_j is taken from the j th stratum, $j = 1, \dots, k$.

2.2 The Simple Linear Homoscedastic Model

This is the usual regression model in which

$$\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon}, \quad (2.1)$$

where β is a $p \times 1$ vector of coefficients, and $\tilde{\epsilon}$ is random error with mean 0 and variance σ^2 . The key assumption is that the mean and variance of $\tilde{\epsilon}$, conditional on (\tilde{X}, \tilde{J}) , are independent of (\tilde{X}, \tilde{J}) .

2.3 The Mixture Model

This model supposes no unique β , but that β varies by stratum in the population. That is, there are k parameter vectors $\beta(1), \dots, \beta(k)$, and, conditional on $\tilde{J} = j$,

$$\tilde{Y} = \tilde{X}\beta(j) + \tilde{\epsilon}, \quad (2.2)$$

where, again, $\tilde{\epsilon}$ has mean 0 and variance σ^2 , independent

of (\tilde{X}, \tilde{J}) . By analogy to the univariate sample survey problem of estimating a mean, one target quantity of interest is the weighted average coefficient, namely, the vector

$$\begin{aligned} \tilde{\beta} &= \sum_{j=1}^k \pi_j \beta(j), \\ &= \sum_{i=1}^n w_i \beta_i / \sum_{i=1}^n w_i, \end{aligned} \quad (2.3)$$

where $\beta_i = \beta(j_i)$ and the second equality follows from (1.2).

2.4 The Omitted-Predictor Model

This model assumes that the simple homoscedastic model of Section 2.2 would hold if only \tilde{X} were augmented by the unfortunately omitted $(1 \times q)$ variable \tilde{Z} . That is, given $(\tilde{X}, \tilde{Z}, \tilde{J})$,

$$\begin{aligned} \tilde{Y} &= \tilde{X}\alpha + \tilde{Z}\gamma + \tilde{\epsilon}, \\ &= \tilde{X}\beta + \tilde{U}\gamma + \tilde{\epsilon}, \end{aligned} \quad (2.4)$$

where $\tilde{\epsilon}$ has mean 0 and variance σ^2 , independent of $(\tilde{X}, \tilde{Z}, \tilde{J})$. The coefficients of \tilde{X} and \tilde{Z} are α and γ , respectively, while \tilde{U} is the part of \tilde{Z} orthogonal to \tilde{X} , namely

$$\tilde{U} = \tilde{Z} - \tilde{X}E(\tilde{X}'\tilde{X})^{-1}E(\tilde{X}'\tilde{Z}). \quad (2.5)$$

Since \tilde{Z} has not been identified, the target quantity or parameter of interest in this model is β , but if \tilde{Z} were identified, we assume the analyst would prefer to know (α, γ) rather than to know merely β .

There are two important points relating this model and the mixture model. First, even if \tilde{Z} is taken to be the $\tilde{X} \times \tilde{J}$ interaction variable, so that the two models are identical, the two parameters β and $\tilde{\beta}$ usually will not be equal. Second, even when assuming that omitted predictors exist, we have in mind that they are not too numerous, so that although the omitted-predictor model is theoretically a generalization of the mixture model, in practice it would have fewer parameters since one would hope that $(p + q) \ll kp$, especially if k , the number of strata, is large.

2.5 The General Nonlinear Model (No Model)

This model makes the minimal assumption that

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\epsilon}^*, \quad (2.6)$$

where $E(\tilde{\epsilon}^*) = 0$ and $\text{cov}(\tilde{X}, \tilde{\epsilon}^*) = 0$. However, no other assumptions are made about $E(\tilde{\epsilon}^* | \tilde{X}, \tilde{J})$ or $V(\tilde{\epsilon}^* | \tilde{X}, \tilde{J})$. The parameter β^* is thus defined as

$$\beta^* = E(\tilde{X}'\tilde{X})^{-1}E(\tilde{X}'\tilde{Y}). \quad (2.7)$$

The parameter β^* will be called the census coefficient, since it would be the least squares estimate if the population were finite and the entire population were sampled, as in a census. Another interpretation of β^* is that $\tilde{X}\beta^*$ represents the best linear predictor of \tilde{Y} in the sense of minimizing the expected squared error of prediction. This

“model” is not really a model since it assumes no population structure except that necessary to define the target quantity β^* .

If every n_j is small compared to the size of the j th population stratum, this model seems equivalent to the finite population formulation in which the values of \bar{Y} and \bar{X} in the population are treated as fixed with no underlying structure. This model includes the three earlier models as special cases. Note, however, that if the mixture model is true, it is *not* generally true that $\beta^* = \bar{\beta}$, while, if the simple homoscedastic model or the omitted predictor model is true, then $\beta^* = \beta$. In fact, setting $\bar{\epsilon}^* = \bar{U}\gamma + \bar{\epsilon}$ shows that the omitted-predictor model is formally equivalent to the general nonlinear model. But the former model assumes that \bar{U} (actually \bar{Z}) is an easily interpreted and not-too-hard to measure variable that was omitted by oversight or by some practical necessity, while the latter model allows \bar{U} to be any variable with $\text{cov}(\bar{U}, \bar{X}) = 0$, perhaps an unobservable variable.

One reason for introducing both models is, as discussed in Section 3.4, to contrast two approaches a statistician might take after rejecting the simple homoscedastic model. The optimist (“a good fitting model is just around the corner”) searches for extra predictors. The pessimist (“models are never valid in the real world”) refuses to rely on any population structure.

3. WHEN TO USE WEIGHTED REGRESSION

3.1 Not If the Simple Linear Homoscedastic Model is Acceptable

Under the linear homoscedastic model, $\hat{\beta}$ is unbiased and has minimum variance among all linear unbiased estimators; it would naturally be preferred to $\hat{\beta}_w$. Haberman (1975) proves various relations between $\hat{\beta}$ and $\hat{\beta}_w$. For example, he shows that for any linear combination c' of the coefficients

$$4R/(1 + R)^2 \leq V(c'\hat{\beta} | J)/V(c'\hat{\beta}_w | J) \leq 1,$$

where R is the ratio of the largest to the smallest of the $\{w_i\}$. In order for the linear homoscedastic model to be acceptable, it must be a priori plausible substantively and in addition pass the usual data analytic tests involving examination of residuals, checking for interactions, and so on.

3.2 Not If the Parameter $\bar{\beta}$ in the Mixture Model is to be Estimated

The mixture model cannot provide a general rationale for preferring $\hat{\beta}_w$ to $\hat{\beta}$ as an estimator of $\bar{\beta}$. To see this, consider the model of Section 2.3 and let $v_{n \times 1}$ and $\mu_{n \times 1}$ be defined by

$$v_i = x_i \beta_i \quad i = 1, \dots, n;$$

$$\mu = X\beta,$$

where x_i is the i th row of X . Then elementary calculations (let $Y = v + \epsilon = \mu + v - \mu + \epsilon$ in (1.1) and (1.3)) show

that

$$E(\hat{\beta} | J) = \bar{\beta} + (X'X)^{-1}X'(v - \mu),$$

$$E(\hat{\beta}_w | J) = \bar{\beta} + (X'WX)^{-1}X'W(v - \mu).$$

Notice that, in general, neither $\hat{\beta}$ nor $\hat{\beta}_w$ is an unbiased estimate of the average coefficient $\bar{\beta}$, and there is no simple way to tell from the two preceding expressions which has the smaller bias. For example, if $p = 1$, so that $\bar{\beta}$, the $\hat{\beta}_i$, and the x_i are all scalars, then

$$\text{Bias}(\hat{\beta}) = \sum x_i^2(\beta_i - \bar{\beta})/\sum x_i^2,$$

$$\text{Bias}(\hat{\beta}_w) = \sum w_i x_i^2(\beta_i - \bar{\beta})/\sum w_i x_i^2.$$

Then, if $x_i \equiv 1$, $\text{Bias}(\hat{\beta}_w) = 0$ by the definition (2.3) of $\bar{\beta}$, but for other choices of X this will not be true. In fact, it could happen that x_i^2 is proportional to w_i , in which case $\hat{\beta}$ would be unbiased but $\hat{\beta}_w$ would not be. In general, neither $\hat{\beta}$ nor $\hat{\beta}_w$ appear to be suitable estimators for $\bar{\beta}$ in the mixture model. Konijn (1962) and Porter (1973) use the mixture model and recommend estimating β separately within each stratum and then taking a weighted average of the estimates as the final estimate of $\bar{\beta}$. That is, use

$$\hat{\beta} = \sum \pi_j \hat{\beta}(j) = \sum w_i \hat{\beta}_i / \sum w_i.$$

Unfortunately, this recommendation is inadvisable for sampling schemes with many strata and relatively few observations per stratum. Pfefferman and Nathan (1981) suggest using weights for the $\hat{\beta}_i$ that take into account the precision of each $\hat{\beta}_i$. Sometimes separate estimation within many strata is impossible because there are too few degrees of freedom. If one were especially suspicious that one of the coefficients (typically the constant term of the usual regression) varies by strata, one could allow the estimate of only that coefficient to vary by strata. Such an analysis of covariance on the entire data set costs only one degree of freedom per stratum.

3.3 Use $\hat{\beta}_w$ If the Linear Homoscedastic Model is not Acceptable but an Estimate of β^* is Desired

The advantage of $\hat{\beta}_w$ in the models of Sections 2.4 and 2.5 is that $\hat{\beta}_w$ is at least a consistent estimator of $\beta = \beta^*$, while $\hat{\beta}$ may not be. Proof of consistency: let each $n_j \rightarrow \infty$ and $w_i \propto \pi_{ji}/n_{ji}$, $i = 1, \dots, n$. Then with probability one $X'WX/\sum w_i$ approaches $E(\bar{X}'\bar{X})$ and $X'WY/\sum w_i$ approaches $E(\bar{X}'\bar{Y})$ so that by (2.7) $\hat{\beta}_w \rightarrow \beta^*$. On the other hand, if the sample sizes of the strata, n_j , are not proportional to the population proportions π_j (i.e., the w_i do not approach equality), then $\hat{\beta}$ need not approach β^* .

3.4 A Strategy for Choosing Between $\hat{\beta}$ and $\hat{\beta}_w$

First, if one believes the mixture model of Section 2.3 and desires to estimate $\bar{\beta}$ of Equation (2.3), then neither $\hat{\beta}$ nor $\hat{\beta}_w$ is appropriate. Therefore, for the rest of this article we ignore the mixture model and the estimation of $\bar{\beta}$.

There remains the problem of choosing between the linear homoscedastic model of Section 2.2 (thus choosing $\hat{\beta}$) and the more general models of Sections 2.4 and 2.5. If one prefers to estimate β^* within the general nonlinear model, $\hat{\beta}_w$ is appropriate. If one believes the omitted-predictor model, then one should try to identify the extra predictor Z and estimate (α, γ) in Equation (2.4) or, failing that, settle for using $\hat{\beta}_w$ as an estimate of β .

The controversy arises in deciding how much evidence, if any, to require before giving up the linear homoscedastic model. Closely related is the question of how hard to look for additional predictors. On one side are those (see Kish and Frankel 1974, Brewer and Mellor 1973, and references therein) who tend to be extremely dubious of the assumptions of the linear homoscedastic model and who also may not be very interested in searching for extra predictors. They are satisfied with making inferences about the census parameter β^* .

On the other side are those who tend to accept the simple model of Section 2.2 so long as it can withstand the scrutiny of a careful regression analysis as described, for example, in the books by Mosteller and Tukey (1977) and Belsley, Kuh, and Welsch (1980). The process of refitting with transformed variables, checking for interactions, plotting residuals, and so on, may lead to the use of extra predictors, but the basic strategy is to accept the simple model (and use $\hat{\beta}$) unless evidence against it develops. The advantages of this approach are, one, the simple inferential procedures (standard F tests and confidence intervals) and, two, the more straightforward interpretation of β , which the model of Section 2.2 allows. Without the assumptions of that model, the interpretation of β^* is difficult. For example, years of schooling may have a positive β^* for predicting income, but the income of some subgroups may drop with increasing education. Published regression analyses are often applied to subpopulations or to completely different populations by later researchers. In that case, β^* may be misleading, while the extra effort spent to identify interactions or other omitted predictors may lead to greater theoretical understanding. Smith (1976) makes a similar point.

Another way to contrast the opposing sides of the controversy is by the priority they assign to defining a target quantity. One side first defines the target quantity β^* as the parameter of interest, no matter what the population structure may be. The other side tends to spend more effort searching for structure in the population and then chooses, from among target quantities like β , $\hat{\beta}$, (α, γ) , and β^* that parameter most suited to whatever structure seems to be present.

In the spirit of this latter approach, we next describe yet another test that the data should pass before one accepts the simple model and uses the estimator $\hat{\beta}$.

4. USING THE WEIGHTS TO TEST THE SIMPLE MODEL

The test is based on the difference $\hat{\Delta} = \hat{\beta}_w - \hat{\beta}$. The assumptions of Section 2.2 imply that $\Delta \equiv E(\hat{\Delta}) = 0$. As an alternative hypothesis, we consider the omitted-pre-

dictor model of Section 2.4, $Y = X\alpha + Z\gamma + \epsilon$, where the columns of Z are further (perhaps unobserved) predictors that should have been included in the regression but were not. Although this particular alternative hypothesis formulation is not essential here, its use will be convenient for expressing and interpreting the expected mean squares of an ANOVA test of $\Delta = 0$ using standard linear models theory. We will see that the hypothesis $\gamma = 0$ implies $\Delta = 0$ but not vice versa. The hypothesis $\Delta = 0$ can also be interpreted as $E(\hat{\beta} | J) = \beta^*$ in the context of the general model of Section 2.5, but our development concentrates on the use of $\hat{\Delta}$ in a test (perhaps one of many) of the simple model versus the omitted-predictor model. Furthermore, when our test rejects the simple model, examination of $\hat{\Delta}$ usually suggests candidates for the needed predictors Z . In this section we do not distinguish between $E(\cdot)$ and $E(\cdot | J)$, since all expectations here are conditional on (X, Z) and, for the two models being compared, the additional conditioning on J makes no difference.

Since $\hat{\Delta} = \hat{\beta}_w - \hat{\beta}$, it may be represented as $\hat{\Delta} = DY$, where

$$D = (X'WX)^{-1}X'W - (X'X)^{-1}X'.$$

Under the model of Section 2.2, elementary calculations show that the covariance matrices of $\hat{\beta}$, $\hat{\beta}_w$, and $\hat{\Delta}$ are

$$\begin{aligned} V(\hat{\beta}) &= (X'X)^{-1}\sigma^2, \\ V(\hat{\beta}_w) &= (X'WX)^{-1}(X'W^2X)(X'WX)^{-1}\sigma^2, \\ V(\hat{\Delta}) &= DD'\sigma^2 \\ &= [(X'WX)^{-1}(X'W^2X)(X'WX)^{-1} - (X'X)^{-1}]\sigma^2 \\ &= V(\hat{\beta}_w) - V(\hat{\beta}). \end{aligned} \quad (4.3)$$

Notice two things about these expressions. First, $V(\hat{\beta}_w)$ is *not* $(X'WX)^{-1}\sigma^2$, as would be true if $V(\epsilon_i) = \sigma^2/w_i$, $i = 1, \dots, n$. Thus, the standard errors and t statistics output by most weighted regression computer programs are invalid for our situation, even if the linear model holds and $\Delta = 0$. Second, since $V(\hat{\beta}_w) = V(\hat{\beta} + \hat{\Delta}) = V(\hat{\beta}) + V(\hat{\Delta})$, we see that $\hat{\beta}$ and $\hat{\Delta}$ are uncorrelated, as can be shown directly by noticing that, as a linear transformation of Y , $\hat{\Delta}$ is orthogonal to the columns of X (i.e., $D'X = 0$). Therefore, the sum of the squared residuals from the unweighted regression can be partitioned into a part due to Δ and an error, or unexplained, component. (Assume $n > 2p$ and that both $(X'X)$ and $V_\Delta = V(\hat{\Delta})/\sigma^2$ are nonsingular.) This leads to an ANOVA table with three independent components (see Table 1).

If the model of Section 2.2 is true, and in addition ϵ is normally distributed, then the ratio $MS_w/\hat{\sigma}^2$ has an F distribution with p and $(n - 2p)$ degrees of freedom. Under the extended model of Section 2.4, the expected value of MS_w is $\sigma^2 + \Delta'V_\Delta^{-1}\Delta/p$, while the expected value of $\hat{\sigma}^2$ is $\sigma^2 + \tau^2/(n - 2p)$, where

$$\tau^2 = \gamma'Z'(I - X(X'X)^{-1}X')Z\gamma - \Delta'V_\Delta\Delta.$$

The formula for τ^2 can be interpreted as the difference

Table 1. Formulas for ANOVA Table Comparing Weighted and Unweighted Regressions

Source	df	Sum of Squares	Mean Square
Regression ^a	p	$SS_R = \hat{\beta}'(X'X)\hat{\beta}$	$MS_R = SS_R/p$
Weights	p	$SS_W = \hat{\Delta}'V_{\Delta}^{-1}\hat{\Delta}$	$MS_W = SS_W/p$
Error	$n - 2p$	$SS_E = \text{remainder}$	$\hat{\sigma}^2 = SS_E/(n - 2p)$
Total	n	$Y'Y$	

^a The source labeled regression here includes the constant term if it is present in the model. In most applications, and in our examples of Section 7, the effect of the grand mean is omitted and the df for regression is $p-1$, while SS_R and $Y'Y$ are reduced by the square of the grand mean.

between the excess in the residual sum of squares due to neglecting the term $Z\gamma$ in the model (4.1) and that accounted for by estimating $\Delta = DZ\gamma$, where D is given by (4.2). If τ^2 is small (the next theorem implies that $\tau^2 = 0$ is equivalent to $Z = WXC$ for some matrix C), then the F test based on $MS_W/\hat{\sigma}^2$ will be a useful test of the simple linear model of Section 2.2. If this model is rejected, we conclude that $\hat{\beta}$ and $\hat{\beta}_W$ have different expectations. The rationale for preferring unweighted to weighted regression is also rejected unless some other variables Z can be found that lead one to accept an extended model of the form (4.1).

A weighted least squares computer program is required to compute $\hat{\beta}_W$ and $\hat{\Delta}$, and another special program is required to compute V_{Δ} . However, as the following theorem shows, SS_W can be computed and the test performed with ordinary, unweighted, regression programs.

Theorem. The F test for $\Delta = 0$ is the same as the usual F test for $\gamma = 0$, if the regression model $Y = X\alpha + WX\gamma + \epsilon$ is fitted by ordinary least squares. (That is, create the new variables $Z = WX$ and test for the effect of Z partialled on X .)

Proof. Since $\hat{\Delta} = DY = D(X\alpha + WX\gamma + \epsilon)$, and $DX = 0$,

$$\Delta = E(\hat{\Delta}) = DWX\gamma = V_{\Delta}(XWX)\gamma$$

using (4.2) and (4.3), where $V_{\Delta} = V(\hat{\Delta})/\sigma^2$. Therefore, the F tests of $\Delta = 0$ and $\gamma = 0$ will be equivalent if V_{Δ} and (XWX) are both nonsingular. One can show that this condition is equivalent to the assumption that the matrix $(X:WX)$ is full rank ($=2p$), which is true if there are at least $(p+1)$ distinct w_i whose corresponding rows of X have rank p . (We conjecture that the theorem is true for arbitrary X and W , although the F tests would have fewer degrees of freedom in the numerator.)

In practice, one might use one of two different methods to compute SS_R and SS_W , depending on the details of one's least squares regression computer program, after having formed the variables Y , X , and $Z = WX$ ($Z_{ij} = w_i X_{ij}$).

Method A. Perform the regression of Y on X and Z , and then refit the regression, dropping the Z variables.

The two "due to regression" sums of squares will be $SS_R + SS_W$, and SS_R , respectively.

Method B. Perform the regressions of Y and Z on X . Then perform the regression of Y on the residuals of the Z -on- X regressions. The last "due to regression" sum of squares will be SS_W .

5. SOME FURTHER REMARKS

The following remarks are somewhat independent of each other but are offered as discussion and for clarification.

Remark 1. Although the tests involving $\hat{\Delta}$ and $\hat{\gamma}$ are equivalent, interpretation of their individual components is somewhat different and less straightforward for $\hat{\gamma}$ than for $\hat{\Delta}$. If the hypothesis $\Delta = 0$ is rejected, we suggest checking for interactions among the variables for which the corresponding components of $\hat{\Delta}$ or $\hat{\gamma}$ are significantly different from zero. This procedure has, of course, no power to detect interactions not related to the weight variable, that is, those for which the τ^2 of Section 4 is large.

Remark 2. Bishop (1977) suggests using a weighted regression in certain situations when the sampling ratio is a function of the dependent variable (not merely the predictor variables) even when the simple model holds. The present article does not discuss that situation, which is akin to retrospective or case-control sampling. Manski and McFadden (1980) provide a general formulation and analysis of the problem. Holt, Smith, and Winter (1980) provide a different formulation and solution that assumes that the probability of selection depends on a normally distributed design variable with known variance.

Remark 3. Thomson (1978) presents another rationale for using estimates of regression coefficients that depend on the sample design, even when the additive model holds. He shows that although $\hat{\beta}$ is best conditional on X , there may be other unbiased estimators with smaller variance for certain ranges of the true β , if bias and variance are computed by averaging over all values of X in a given sampling design. However, under Thompson's model, $\hat{\beta}$ is always more efficient than $\hat{\beta}_W$.

6. AN EXAMPLE

To elucidate the issues in terms of the applicability of the models discussed in Section 2, we consider the analysis of a subset of data from the Panel Study of Income Dynamics, a continuing longitudinal study begun in 1968 by the University of Michigan's Survey Research Center. The original sample of 4,802 families was composed of two subsamples. The larger portion (2,930) of the original interviews were conducted with household heads from a representative cross-section sample of families in the United States. An additional 1,872 interviews were conducted with heads of low-income households drawn from a sample identified and interviewed by the Census Bureau for the 1966–1967 Survey of Economic Opportunity. Annual interviews have been conducted since 1968 with

these household heads and also with the heads of new families formed by original panel members who have left home. At the end of the fifth year of the study, a set of weights was calculated to account for initial variations in sampling rates and variations in nonresponse rates. The weights were intended to help estimate population means and totals and for possible use in other statistical analyses. The calculation of the weights, which are inversely proportional to the probability of selection for each individual, is described in Morgan (1972, pp. 33–34). For the purpose of this example, we ignore the fact that the sampling scheme was clustered as well as stratified.

In this analysis we attempt to predict educational attainment and we restrict ourselves to panel individuals (a) who were children in 1968 households, aged 14–18; (b) who had become heads or wives of families by 1975; (c) who had completed their schooling by 1975; and (d) who had completed at least eight years of schooling. Restrictions (c) and (d) eliminated 46 and 9 observations, respectively. A final restriction was necessary because the 867 individuals satisfying restrictions (a) through (d) came from only 658 different families. Since the educational attainments of siblings are not likely to be independent, we randomly selected one observation from each set of siblings that came from the same family, reducing the number of observations to 658. The weights for these 658 cases range from 1 to 83, with a mean of 29.0 and standard deviation equal to 21.4. The data used in this analysis, consisting of 867 computer card images, is available from the authors.

Theoretical and empirical studies of the economics of educational attainment (Becker 1975, Ben-Porath 1967, Duncan 1974, Edwards 1975, Hill 1979, Liebowitz 1974, Parsons 1975, and Wachtel 1975) have identified numerous characteristics of the family and the economic environment that may affect the attainment decision. This past research leads to our initial specification of the following form:

$$\begin{aligned}
 Ed = & \alpha + \beta_1 FaEd + \beta_2 MothEd + \beta_3 Sibs \\
 & (+) \quad (+) \quad (-) \\
 & + \beta_4 Family\ Income + \beta_5 Age + \beta_6 Exp/Pupil \\
 & (+) \quad (+) \quad (+) \\
 & + \beta_7 Unemployment + \beta_8 Rural + \beta_9 \%College \\
 & (+) \quad (-) \quad (+) \\
 & + \beta_{10} County\ Income. \\
 & (+)
 \end{aligned} \quad (6.1)$$

Table 2 defines the variables in the model (6.1) and the hypothesized signs of the coefficients are given in the parentheses of (6.1). In addition, we include dummy variables for black males, white females, and black females in order to compare their educational attainment with white males. Because low-income families were over-sampled, the number of black and white observations are far more evenly distributed in the sample than in the population. There are 176 observations on white males, 112

Table 2. Variable Definitions in Model Equation (6.1)

<i>Ed</i>	Completed educational attainment of the individual, in years, self-reported
<i>FaEd</i>	Educational attainment level of the father, in years, reported by the father
<i>MothEd</i>	Educational attainment level of the mother, in years, reported by the father
<i>Sibs</i>	Number of siblings
<i>Family Income</i>	1967–1971 average total parental family income, in thousands of 1967 dollars, reported by the father
<i>Age</i>	The age of the individual, in years
<i>Exp/Pupil</i>	Per student public school expenditure in 1968, for county of residence in 1968
<i>Unemployment</i>	The percent of county labor force unemployed in 1970, for county of residence in 1968
<i>Rural</i>	A dichotomous variable equal to one if the parental family resides more than 50 miles from a city of 50,000 or more in 1968, and zero otherwise
<i>% college</i>	The percent of persons 25 or more years old in the 1968 county of residence who have completed four or more years of college
<i>County Income</i>	The median 1969 income in 1968 county of residence, in thousands of 1969 dollars

on black males, 221 on white females, and 149 on black females.

7. RESULTS OF THE DATA ANALYSES

We began our analysis by obtaining unweighted estimates of the parameters of (6.1), with the race-sex dummy variables included as additive predictors. The coefficients and associated standard errors are the boldface entries in the second and third columns, respectively, of Table 3. Considered as a whole, these variables account for more than a quarter of the variance in educational attainment. Virtually all of them have the hypothesized signs, although, with the exception of the county income variable, only the ones measured at the family level are statistically significant at conventional levels. The results for the county income variable is puzzling, although a significant negative coefficient has also been found by Wachtel (1975, p. 515) with different data. A histogram of the residuals and a scatterplot of the residuals versus the fitted values showed no gross deviations from the assumptions of the simple model.

Using Method A of Section 4 to compare $\hat{\beta}$ with $\hat{\beta}_w$, we formed 14 *Z* variables by multiplying each independent variable (including the constant) by the weight variable. When the dependent variable is regressed on both sets of independent variables, we found that the null hypothesis, that $\Delta = 0$ (i.e., the simple model is correct), could be rejected at about the 6 percent level, as Table 4 indicates. We proceeded to calculate estimates of Δ , V_Δ , and weighted estimates of β .

The weighted estimates of β and the *t* ratios of Δ are given in the boldface entries of columns 4 and 5 of Table 3. The regression performed in Method A provides an estimate of coefficients (γ) and standard errors for each

Table 3. Coefficients, Standard Errors, and *t* Ratios of Various Tests for Two Versions of the Educational Attainment Model

Independent Variable	$\hat{\beta}$: Unweighted Estimate of β		SE($\hat{\beta}$): Standard Error		$\hat{\beta}_w$: Weighted Estimate of β		Δ : <i>t</i> Ratio of Difference Between Weighted and Unweighted Estimates of β		t_y : <i>t</i> Ratio of Estimated Coefficient of Variable \times Weight	
EaEd	.082	.076	(.021)	(.020)	.082	.082	.0	.3	.0	.1
MothEd	.125	.002	(.026)	(.126)	.158	.055	1.6	.6	2.0	.5
Sibs	-.073	-.069	(.031)	(.031)	-.069	-.074	.2	-.2	.0	-.2
Family Income	-.044	.033	(.012)	(.012)	.039	.030	-.9	-.5	-.2	-.5
Age	.271	.251	(.045)	(.045)	.314	.297	1.3	1.4	1.6	1.6
Whether Black Male	-.280	-1.267	(.215)	(.448)	-.639	-1.753	-1.5	-1.0	-2.1	-2.0
Whether White Female	-.068	1.385	(.163)	(.796)	-.218	1.048	-2.8	-.5	-2.8	-1.1
Whether Black Female	-.070	.491	(.196)	(.900)	.142	.268	1.0	-.3	-.3	1.0
Expenditure/Pupil	.022	.006	(.078)	(.077)	.027	.027	.1	.4	.5	.6
Unemployment	.037	.183	(.042)	(.076)	-.015	.210	-2.3	.3	-2.3	.3
Rural	-.081	-.097	(.175)	(.172)	-.113	-.119	-.3	-.2	.2	.2
% College	.016	.019	(.017)	(.017)	.030	.024	1.0	.4	.6	.4
County Income	-.104	-.084	(.049)	(.048)	-.129	-.092	-.7	-.2	-.7	-.3
Constant	6.705	7.679	(.907)	(1.080)	6.073	6.794	-1.0	-1.1	-.7	-.6
Whether White and Unemployment		-.208		(.085)		-.252		-.5		-1.2
Whether Female and MothEd		-.340		(.169)		-.320		.2		.2
MothEd ²		.008		(.007)		.005		-.6		-.4
Whether Female \times MothEd ²		.017		(.009)		.017		-.0		-.2
R ²	.289	.315								
Standard Error of Estimate	1.60	1.57								
Sample Size	658	658								

Source: Morgan (1972).

of the *Z* variables; the *t* ratios of each are given in the sixth column of the table.

Given the significance of these differences, we could have chosen to use the $\hat{\beta}_w$ of column 4 as descriptive estimates of β^* in the census model. This rejection of the simple model would put more emphasis on race and sex, and less on unemployment, as predictors of educational attainment.

Table 4. ANOVA Table Comparing Weighted and Unweighted Regressions Using the Initial Model

Source	df	Sum of Squares	Mean Square	F	Significance
Regression	13	670.0	51.5	20.1	<.0001
Weights	14	59.2	4.2	1.7	.056
Error	630	1586.7	2.5		
Total	657	2315.9			

We chose instead to use the information in the boldface entries of Table 3 to explore extensions of the simple model of (6.1). The unweighted coefficients differed substantially from the weighted coefficients for four variables: mother's education, county unemployment rate, and two of the race-sex dummy variables. Notice that there is a rough correspondence between the ranking (and direction) of the *t* ratios on differences between unweighted and weighted coefficients and the *t* ratios of the corresponding *Z* variables. Thus, the more easily computed tests of significance on the *Z* variables can apparently serve as a guide to variables with large Δ 's.

Prior research and the significant Δ 's suggest that the most probable cause of misspecification is omitted interactions between race and sex and some of the independent variables listed in (6.1), especially mother's education and county unemployment. Although the hypothesis of equal slopes for the four race and sex subgroups could not be rejected ($F = .94$; $df = 30, 614$),

the subset regressions did suggest a possible interaction between race and county unemployment rate in which increases in unemployment had a stronger positive effect on the educational attainment of blacks than of whites. This interaction is quite plausible since unemployment rates for blacks are considerably higher than those of whites and a unit change in the *overall* county unemployment rate has more effect on blacks than on whites. When this interaction term was added to (6.1), the coefficient was $-.21$ with a standard error of $.09$. Furthermore, when Method A was repeated with it and its associated Z variable (i.e., whether white \times county unemployment rate \times weight) included as predictors, the F ratio of the entire set of weight interactions drops from 1.7 to 1.3, and the coefficients of the two Z variables formed from the county unemployment variable are insignificant.

Since the three other Z variables significant at the $.05$ level in the original specification remained significant when the race-unemployment interaction was included, we continued our search for additional interactions. We discovered that mother's education interacted with itself (i.e., its effect was nonlinear) and, furthermore, that these nonlinear effects of mother's education on the educational attainment of the children depended upon the sex of the child.

Results from the estimation of our final specification of the education attainment model are the italicized entries of Table 3. In contrast to the initial specifications, the highest t ratio for the difference between weighted and unweighted coefficients is 1.4. The analysis of variance table for the final model, Table 5, shows that the F ratio associated with the weights is below 1.0.

The coefficients and standard errors presented as the italic entries of columns 2 and 3 of Table 3 should be regarded with some caution because the data were used to suggest the appropriate functional form.

As a final analysis step, we reestimated the model on the 209 individuals who were excluded when the sample was restricted to only one sibling per family. The results were quite similar, particularly for the race-unemployment interaction and the nonlinear effect of mother's education. Although the sign of the interaction between sex of child and mother's education changed direction, the only coefficient that changed by a statistically significant amount was family income. The income coefficient increased, suggesting a more important role for income for

Table 6. Estimated Increase in Mean Ed per Year of Increase in MothEd, Based on Final Model

Sex of Child	Education Level of Mother		
	8 Grades	12 Grades (H. S. Grad.)	16 Grades (College Grad.)
Male	.13	.19	.26
Female	.06	.26	.46

these 209 individuals from families with at least two children in this five-year age cohort. When the final model was fitted to the complete sample of 867 individuals, the coefficient of income rose from $.033$ to $.044$.

Our use of the weights to test for model misspecification has led us to the substantive conclusion that a simple linear additive educational attainment model is not appropriate for several reasons. First, a worsening of local economic conditions (as measured by the county unemployment rate) appears to provide more of an incentive for blacks to stay in school than for whites. A one percentage point increase in the unemployment rate was associated with an additional one-fifth of a year of educational attainment for blacks, while the effect for whites was essentially zero. Second, the effects of mother's education on the educational attainment of children increase with the level of her education and furthermore depend on the sex of the child. Table 6 evaluates $\partial \text{Ed} / \partial \text{MothEd}$ for mothers with 8, 12, and 16 years of education.

There is a modest increase in the marginal effect of mother's educational attainment for sons and a much more dramatic increase in this effect for daughters. An extra year of mother's education level is associated with virtually no increase in the attainment of daughters whose mothers have an eighth grade education but is associated with an additional one-half year of education for daughters with college-educated mothers. These conclusions do not change when the model is estimated from the complete sample of 867 individuals. None of the six numbers in this table changes by more than $.03$.

Finally, an analysis of the residuals from this extended model revealed no anomalies, and we thus prefer the unweighted estimates of its coefficients.

[Received November 1976. Revised November 1982.]

REFERENCES

- BACHMAN, J.G., GREEN, S., and WIRTANEN, I.D. (1974), *Youth in Transition, Vol. III, Dropping Out—Problem or Symptom?*, Ann Arbor, Mich.: Institute for Social Research.
- BECKER, GARY S. (1975), *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* (2nd ed.), New York: Columbia University Press.
- BELSLEY, D.A., KUH, E., and WELSCH, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley.
- BEN-PORATH, YORAM (1967), "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, 75, 352–365.
- BISHOP, JOHN (1977), "Estimation When the Sampling Ratio is a

Table 5. ANOVA Table Comparing Weighted and Unweighted Regressions Using the Final Model

Source	df	Sum of Squares	Mean Square	F	Significance
Regression	17	730.6	43.0	17.35	<.0001
Weights	18	43.3	2.4	.97	.494
Error	622	1542.0	2.5		
Total	657	2315.9			

- Linear Function of the Dependent Variable," *Proceedings of Social Statistics Section of the American Statistical Association*, 848–853.
- BLUMENTHAL, M.D., KAHN, R.L., ANDREWS, F.M., and HEAD, K.B. (1972), *Justifying Violence: Attitudes of American Men*, Ann Arbor, Mich.: Institute for Social Research.
- BREWER, K.R.W., and MELLOR, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys," *Australian Journal of Statistics*, 15, 145–152.
- DRAPER, N.R., and SMITH, H. (1966), *Applied Regression Analysis*, New York: John Wiley.
- DUNCAN, GREG (1974), "Educational Attainment," *Five Thousand American Families—Patterns of Economic Progress*, Vol. I, eds. Morgan et al., Ann Arbor, Mich.: Institute for Social Research, 305–332.
- DUNCAN, GREG J., and MORGAN, JAMES N. (eds.) (1976), *Five Thousand American Families—Patterns of Economic Progress*, Vol. IV, Ann Arbor, Mich.: Institute for Social Research.
- EDWARDS, LINDA N. (1975), "The Economics of Schooling Decisions: Teenage Enrollment Rates," *Journal of Human Resources*, 10, 155–73.
- HABERMAN, SHELBY J. (1975), "How Much Do Gauss-Markov and Least-Square Estimates Differ? A Coordinate-Free Approach," *The Annals of Statistics*, 3, 982–990.
- HILL, C. RUSSELL (1979), "Capacities, Opportunities, and Educational Investments: The Case of the High School Dropout," *The Review of Economics and Statistics*, 61, 9–20.
- HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980), "Regression Analysis of Data from Complex Surveys," *Journal of the Royal Statistical Society, Ser. A*, 143, 474–487.
- HU, T.W., and STROMSDORFER, E.W. (1970), "A Problem of Weighting Bias in Estimating the Weighted Regression Model," *Proceedings of the Business and Economics Section of the American Statistical Association*, 513–516.
- JUSTER, F. THOMAS (ed.) (1976), *The Economic and Political Impact of General Revenue Sharing*, Washington, D.C.: U.S. Government Printing Office.
- KISH, LESLIE (1965), *Survey Sampling*, New York: John Wiley.
- KISH, LESLIE, and FRANKEL, MARTIN R. (1974), "Inference from Complex Samples," *Journal of Royal Statistical Society, Ser. B*, 36, 1–37.
- KLEIN, L.R. (1953), *A Textbook of Econometrics*, Evanston, Ill.: Row, Peterson, and Co.
- KLEIN, L.R., and MORGAN, JAMES N. (1951), "Results of Alternative Statistical Treatments of Sample Survey Data," *Journal of American Statistical Association*, 46, 442–460.
- KONIJN, H.S. (1962), "Regression Analysis in Sample Surveys," *Journal of American Statistical Association*, 57, 590–606.
- LEIBOWITZ, ARLEEN (1974), "Home Investments in Children," *Journal of Political Economy*, 82, 111–131.
- MANSKI, C., and McFADDEN, D. (1980), "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data*, eds. C. Manski and D. McFadden, Cambridge, Mass.: M.I.T. Press.
- MORGAN, JAMES N. (1972), *A Panel Study of Income Dynamics: Study Design, Procedures, Available Data*, Vol. I, Ann Arbor, Mich.: Institute for Social Research.
- MOSTELLER, F., and TUKEY, J.W. (1977), *Data Analysis and Regression*, Reading, Mass.: Addison-Wesley.
- PARSONS, DONALD O. (1975), "Intergenerational Transfers and the Educational Decisions of Male Youth," *Quarterly Journal of Economics*, 89, 603–617.
- PFEFFERMANN, D., and NATHAN, G. (1981), "Regression Analysis of Data From a Cluster Sample," *Journal of the American Statistical Association*, 76, 681–689.
- PORTER, RICHARD D. (1973), "On the Use of Survey Sample Weights in the Linear Model," *Annals of Economic and Social Measurement*, 212, 141–158.
- SMITH, KENT W. (1976), "Analysing Disproportionately Stratified Samples With Computerized Statistical Packages," *Sociological Methods and Research*, 5, 207–230.
- THOMSEN, IB (1978), "Design and Estimation Problems When Estimating a Regression Coefficient From Survey Data," *Metrika*, 25, 27–36.
- WACHTEL, PAUL (1975), "The Effect of School Quality on Achievement, Attainment Levels, and Lifetime Earnings," *Explorations in Economic Research*, 2, 502–536.