USING SAMPLE SURVEY WEIGHTS IN MULTIPLE REGRESSION ANALYSES OF STRATIFIED SAMPLES

William H. DuMouchel, Massachusetts Institute of Technology
Greg J. Duncan, University of Michigan

Abstract: 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 method leads 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.

1. INTRODUCTION

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

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

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 paper will be concerned only with stratified sampling and not 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 paper 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, ..., k are known. Let n_j be the size of a simple random sample drawn from the jth stratum, j = 1, ..., k, so that $n_1 + \ldots + n_k = n$. Since the jth stratum is underrepresented in the sample by a factor proportional to n_j/π_j, the weight assigned to the ith observation is

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

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

$$\hat{\beta}_W = (X'WX)^{-1}\, X'WY \ . \qquad (1.3)$$

Which estimator should be used? Controversy has raged at least since Klein and Morgan (1951). The advocate 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 which 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 paper is to clarify this issue by showing how the appropriate estimator depends on which of 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. Section 2 defines four different regression models of increasing generality which might be used to justify the use of $\hat{\beta}_W$. Section 3 discusses the relationship between the models and the choice of estimator. Section 4 shows how an easily computed test based on $\hat{\beta}_W - \hat{\beta}$ may help in choosing a model. Section 5 contains further discussion, and the last two sections 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 to use the weights or not depends on what one assumes about the population from which the data has been drawn. This section describes four models which exemplify the most common assumptions.

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 shall always assume that the stratum sample size $\{n_j\}$ are small fractions of the corresponding population stratum sizes, and the mathematics of sampling with replacement or from infinite populations will be used throughout this paper. Let $\tilde{Y}$ and $\tilde{X}$ denote the scalar and (1 × 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(·) refer to a simple random sample from the population, while conditional expectations E(·|J) refer to stratified sampling, where a simple random sample of size n_j is taken from the jth stratum, j = 1,...,k.

2.2 The Simple Linear Homoschedastic Model.

This is the usual regression model in which

$$\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon} , \qquad (2.1)$$

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

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 β(1), ..., β(k), and, conditional on $\mathfrak{J} = j$,

$$\tilde{Y} = \tilde{X}\beta(j) + \tilde{\varepsilon} , \qquad (2.2)$$

where, again, $\tilde{\varepsilon}$ has mean 0, variance σ^2, independent of $(\tilde{X},\mathfrak{J})$. By analogy to the univariate sample survey problem of estimating a mean, one vector parameter of interest is the weighted average coefficient,

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

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 §2.2 would hold if only $\tilde{X}$ were augmented by the unfortunately omitted (1 × q) variable $\tilde{Z}$. That is,

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

where $\tilde{\varepsilon}$ has mean 0, variance σ^2, independent of $(\tilde{X},\tilde{Z},\mathfrak{J})$. 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

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

Since $\tilde{Z}$ has not been identified, the 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 of contrast between this model and the mixture model. First, even if $\tilde{Z}$ is taken to be the $\tilde{X} \times \mathfrak{J}$ interaction variable, so that the two *models* are identical, the two *parameters* β and $\bar{\beta}$ *will* not usually 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 would hope that (p + q) << 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

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

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

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

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.

If every n_j is small compared to the size of the jth population stratum, this model seems equivalent to the finite population formulation in which the values of $\tilde{Y}$ and $\tilde{X}$ 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 $\tilde{\varepsilon}^* = \tilde{U}\gamma + \tilde{\varepsilon}$ shows that the omitted-predictor model is formally equivalent to the general nonlinear model. But the former model assumes $\tilde{U}$ (actually, $\tilde{Z}$) is an easily interpreted and not-too-hard to measure variable that was omitted by oversight or some practical necessity, while the latter model allows $\tilde{U}$ to be any variable with $\mathrm{Cov}(\tilde{U},\tilde{X}) = 0$, perhaps an unobservable variable.

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, and 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

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

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, etc.

3.2 Not If the Mixture Model is Preferred.

The mixture model cannot provide a general rationale for preferring $\bar{\beta}_W$ to $\bar{\beta}$. To see this, assume the model of §2.3 and let $\nu_{n\times 1}$ and $\mu_{n\times 1}$ be defined by

$$\begin{aligned} \nu_i &= x_i \beta_i \qquad i = 1, \ldots, n ; \\ \mu &= \tilde{X}\beta , \end{aligned}$$

where x_i is the ith row of X. Then elementary calculations (let Y = ν + ε = μ + ν − μ + ε in (1.1) and (1.3)) show that

$$\begin{aligned} E(\hat{\beta}|J) &= \bar{\beta} + (X'X)^{-1}X'(\nu - \mu) , \\ E(\hat{\beta}_W|J) &= \bar{\beta} + (X'WX)^{-1}X'W(\nu - \mu) . \end{aligned}$$

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 above 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:

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

Then, if $x_i \equiv 1$, 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, but $\hat\beta_W$ would not be unbiased. 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

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

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 $\hat\beta_i$ which 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 §2.4, 2.5 is that $\hat\beta_W$ is at least a consistent estimator of β = β*, while $\hat\beta$ may not be. Proof of consistency: let each $n_j \to \infty$ and $w_i \propto \pi_{j_i}/n_{j_i}$, i = 1, ..., n. Then with probability one $X'WX/\sum w_i$ approaches $E(\tilde X'\tilde X)$ and $X'WY/\sum w_i$ approaches $E(\tilde X'\tilde Y)$ so that by (2.7) $\hat\beta_W \to \beta^*$. 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 §2.3 and desires to estimate $\bar\beta$ of eq.(2.3), then neither $\hat\beta$ nor $\hat\beta_W$ is appropriate. Therefore, the rest of this paper will ignore the mixture model and the estimation of $\bar\beta$.

There remains the problem of choosing between the linear homoscedastic model of §2.2 (thus choosing $\hat\beta$) and the more general models of §2.4 and §2.5. If the general nonlinear model is chosen, $\hat\beta_W$ is appropriate. If one believes the omitted-predictor model, then one should try to identify the extra predictor Z and estimate (α,γ) in eq.(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 §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 (1976) and Belsley, Kuh and Welsch (1980). The process of refitting with transformed variables, checking for interactions, plotting residuals, etc., 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 §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.

In the spirit of this latter approach, we next describe yet another test which 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 $\hat\Delta = \hat\beta_W - \hat\beta$, where $\Delta = E(\hat\Delta) = E(\hat\beta_W) - E(\hat\beta)$. The hypotheses of §2.2 imply that Δ = 0. As an alternative hypothesis, we consider the omitted-predictor model of §2.4:

$$Y = X\alpha + Z\gamma + \varepsilon , \qquad (4.1)$$

where the columns of Z are further (perhaps unobserved) predictors which should have been included in the regression but were not. We will see that the hypothesis γ = 0 implies Δ = 0 but not vice versa. The hypothesis Δ = 0 can also be interpreted as $E(\hat\beta|J) = \beta^*$ in the context of the general model of §2.5, but our development will concentrate 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(•) and E(•|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 $D = (X'WX)^{-1}X'W - (X'X)^{-1}X'$. (4.2)

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

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

Notice two things about the above expressions. First, $V(\hat{\beta}_W)$ is not $(X'WX)^{-1}\sigma^2$, as would be true if $V(\varepsilon_i) = \sigma^2/w_i$, $i = 1, \ldots, n$. 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 $V(\hat{\beta}_W) = V(\hat{\beta} + \Delta) = V(\hat{\beta}) + V(\hat{\Delta})$, 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., DX = 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:

Source	df	Sum of Squares	Mean Square
Regression[1]	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 = remainder	$\hat{\sigma}^2 = SS_E/(n-2p)$
	n	$Y'Y$	

If the model of §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 §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 between the excess in the residual sum of squares due to neglecting the tgerm $Z\gamma$ 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/σ^2 will be a useful test of the simple linear model of §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 which 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 $Y = X\alpha + WX\gamma + \varepsilon$ 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 $\hat{\Delta} = DY = D(X\alpha + WX\gamma + \varepsilon)$, and $DX = 0$,

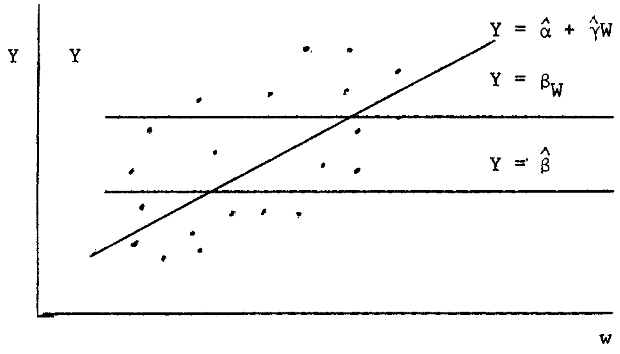
$$\begin{aligned} \Delta = E(\hat{\Delta}) &= DWX\gamma \\ &= V_\Delta(XWX)\gamma \end{aligned}$$

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. This condition can be shown to be 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.)

Figure A shows a geometric interpretation of the theorem when p = 1 and $x_i \equiv 1$ (estimation of the mean of Y only). In this case we can plot the values (w_i, y_i) and draw lines on the plot corresponding to the unweighted mean, weighted mean, and the regression of Y on the weights. Then the theorem states that the test that the two horizontal lines have the same expected value is equivalent to the test for zero slope of the regression line.

Figure A
Geometric Interpretation of the Theorem
when p = 1 and $x_i \equiv 1$.

The test that the two horizontal lines have the same expected value is equivalent to the test that the regression line and the $Y = \hat{\beta}$ line have the same expected value.



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 $Z = WX$ $(Z_{ij} = w_iX_{ij})$.

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.

5.1 Remark One.

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.

5.2 Remark Two.

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 paper 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.

5.3 Remark Three.

Thomson (1978) presents another rationale for using estimates of regression coefficients which 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. A LARGE-SCALE 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 1966-67 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 non-response 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.

This analysis attempts to predict educational attainment and is restricted to panel individuals (1) who were children in 1968 households, aged 14-18; (2) who had become heads or wives of families by 1975; (3) who had completed their schooling by 1975; and (4) who had completed at least eight years of schooling. Restrictions (3) and (4) eliminated 46 and 9 observations, respectively. A final restriction was necessary because the 867 individuals satisfying restrictions (1) through (4) 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 and also from JASA.

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:

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

Table 1 defines the variables in the above model and the hypothesized signs of the coefficients are given in the parentheses of eq.(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.[2]

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 2. Taken 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 result 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 scatter plot of the residuals versus the fitted values, showed no gross deviations from the assumptions of the simple model.

Table 1. Variable Definitions in Model of Equation (6.1)

Ed	Completed educational attainment of the individual, in years, self-reported
FaEd	Educational attainment level of the father, in years, reported by the father
MothEd	Educational attainment level of the mother, in years, reported by the father
Sibs	Number of siblings
Family Income	1967-1971 average total parental family income, in thousands of 1967 dollars, reported by the father
Age	The age of the individual, in years
Exp/Pupil	Per student public school expenditure in 1968, for county of residence in 1968
Unemployment	The percent of county labor force unemployed in 1970, for county of residence in 1968
Rural	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
% College	The percent of persons 25 or more years old in the 1968 county of residence who have completed four or more years of college
County Income	The median 1969 income in 1968 county of residence, in thousands of 1969 dollars

Using Method A to compare $\hat{\beta}$ with $\hat{\beta}_W$, we formed 14 Z variables by multiplyuing 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 the following analysis of variance table indicates:

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		
	657	2315.9			

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 2. The regression performed in Method A provides an estimate of coefficients (γ) and standard errors for each 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.

We chose instead to use the information in the boldface entries of Table 2 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, it would appear that the more easily computed tests of significance on the Z variables can 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 equation (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 ($F = 0.94$; 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 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 whites. When this interaction term was added to equation (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 × county unemployment rate × 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 italic entries of Table 2. 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 2

COEFFICIENTS, STANDARD ERRORS, AND t-RATIOS
OF VARIOUS TESTS FOR TWO VERSIONS OF THE EDUCATIONAL ATTAINMENT MODEL

Boldface: Initial model (eq. (6.1)) *Italic:* Final model chosen as described in text

Independent Variable	$\hat{\beta}$ Unweighted Estimate of β		$SE(\hat{\beta})$: Standard Error		$\hat{\beta}_w$: Weighted Estimate of β		t_Δ: t-Ratio of Difference Between Weighted and Unweighted Estimates of β		t_γ: t-Ratio of Estimated Coefficient of Variable x Weight	
FaEd	**.082**	.076	**(.021)**	(.020)	**.082**	.082	**0.0**	0.3	**0.0**	0.1
MothEd	**.125**	.002	**(.026)**	(.126)	**.158**	.055	**1.6**	0.6	**2.0**	0.5
Sibs	**-.073**	-.069	**(.031)**	(.031)	**-.069**	-.074	**0.2**	-0.2	**0.0**	-0.2
Family Income	**.044**	.033	**(.012)**	(.012)	**.039**	.030	**-0.9**	-0.5	**-0.2**	-0.5
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**	-0.5	**-2.8**	-1.1
Whether Black Female	**-.070**	.491	**(.196)**	(.900)	**.142**	.268	**1.0**	-0.3	**-0.3**	-1.0
Expenditure/ Pupil	**.022**	.006	**(.078)**	(.077)	**.027**	.027	**0.1**	0.4	**0.5**	0.6
Unemployment	**.037**	.183	**(.042)**	(.076)	**-.015**	.210	**-2.3**	0.3	**-2.3**	0.3
Rural	**-.081**	-.097	**(.175)**	(.172)	**-.113**	-.119	**-0.3**	-0.2	**0.2**	0.2
%College	**.016**	.019	**(.017)**	(.017)	**.030**	.024	**1.0**	0.4	**0.6**	0.4
County Income	**-.104**	-.084	**(.049)**	(.048)	**-.129**	-.092	**-0.7**	-0.2	**-0.7**	-0.3
Constant	**6.705**	7.679	**(.907)**	(1.080)	**6.073**	6.794	**-1.0**	-1.1	**-0.7**	-0.6
Whether White and Unemployment		-.208		(.085)		-.252		-0.5		-1.2
Whether Female and MothEd		-.340		(.169)		-.320		0.2		0.2
MothEd^2		.008		(.007)		.005		-0.6		-0.4
Whether Female x MothEd^2		.017		(.009)		.017		-0.0		-0.2
R^2	**.289**	.315								
Standard Error of estimate	**1.60**	1.57								
Sample Size	**658**	6 58								

Source: Morgan (1972)

table for the final model shows that the F-ratio associated with the weights is below 1.0:

Source	df	Sum of Squares	Mean Square	F	Significance
Regression	17	730.6	43.0	17.35	<.0001
Weights	18	43.3	2.4	0.97	.494
Error	622	1542.0	2.5		
	657	2315.9			

The coefficients and standard errors presented as the italic entries of columns 2 and 3 of Table 2 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 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 to stay in school for blacks than 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 attainment of children increase with the level of her education, and furthermore depend upon the sex of the child. The following table evaluates $\partial Ed/\partial MothEd$ for mothers with 8, 12, and 16 years of education:

Education Level of Mother

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

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 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, this extended model seems to fit well, and we thus prefer the unweighted estimates of its coefficients.

8. ACKNOWLEDGEMENTS

We want to thank Professor Herman Chernoff for his careful reading and constructive comments on an earlier draft of this paper. We also want to thank two editors, three associate editors, and four referees for their helpful comments on earlier drafts, and Martin Frankel, Jr., Graham Kalton, and Roger Wright for valuable discussion.

FOOTNOTES

[1] 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.

[2] Because low-income families were oversampled, 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 on black males, 221 on white females and 149 on black females.

REFERENCES

Bachman, Jerald G., et al. (1974) Youth in Transition, Vol. III, Dropping Out--Problem or Symptom? Ann Arbor: 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, 352-65.

Bishop, John (1977) "Estimation When the Sampling Ratio is a Linear Function of the Dependent Variable," Proceedings of Social Statistics Section of the American Statistical Assocation.

Blumenthal, Monica D., et al. (1972) Justifying Violence: Attitudes of American Men. Ann Arbor: 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.

Draper, N.R. and Smith, H. (1966) Applied Regression Analysis. New York: John Wiley.

Duncan, Greg (1974) "Educational Attainment." In James N. Morgan et al., Five Thousand American Families--Patterns of Economic Progress, Vol. I. Ann Arbor: Institute for Social Research, 305-32.

Duncan, Greg J. and Morgan, James N., eds. (1976) Five Thousand American Families--Patterns of Economic Progress, Vol. IV. Ann Arbor: Institute for Social Research.

Edwards, Linda N. (1975) "The Economics of Schooling Decisions: Teenage Enrollment Rates," Journal of Human Resources 1Q, 155-73.

Hausman, Jerry and Wise, David (1977) "Social Experimentation, Translated Distributions, and Efficient Estimation," Econometrica 45: 919-938.

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.

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 et al. (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, Series B, 1-37.

Klein, L.R. (1953) A Textbook of Econometrics. Evanston: 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: 442-60.

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, Part II.

Manski, C. and McFadden, D. (1980) "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in Manski, C. and D. McFadden (eds.), Structural Analysis of Discrete Data. Cambridge: M.I.T. Press.

Morgan, James N. (1972) A Panel Study of Income Dynamics: Study Design, Procedures, Available Data, Vol. I. Ann Arbor: 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: 603-17.

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.

Shah, V.B., Holt, M.M., Folsom, R.E. (1977) "Inference About Regression Models from Sample Survey Data." Research Triangle Park, N.C. 27709.

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: 502-36.