

2000-4

‘MARK : ONE OR MORE RACES...’:

A Simple Method for Statistical Analyses Involving Multiple Racial Identifications

Glenn Deane

Department of Sociology and
Center for Social and Demographic Analysis
University at Albany, State University of New York

Revised
1/03/02

DRAFT COPY B NOT FOR CITATION OR QUOTATION WITHOUT THE AUTHORS-
PERMISSION

**An earlier version of this paper (June, 2000) was entitled, “‘Check all that Apply...’: A Simple Method for Correcting Observational Dependence in Multiple Response Data.” Support for this research was provided by grants to the Center for Social and Demographic Analysis from NICHD (P30 HD32041) and NSF (SBR-9512290). Opinions, findings, and conclusions expressed here are those of the author and do not necessarily reflect the views of the funding agencies. Address all correspondence to Professor Glenn Deane, Department of Sociology, University at Albany, Albany, NY 12222 email: gdd@csc.albany.edu

‘MARK : ONE OR MORE RACES...’:

A Simple Method for Statistical Analyses Involving Multiple Racial Identifications

Abstract

Multiple racial identifications pose a vexing challenge to individual-level analyses. Typically, data analysts make arbitrary decisions to allocate multi-racial respondents to single-race categories. Treated in this manner, race may be employed as a nominal-level covariate of any number of interesting and important social, economic, or demographic dependent variables, but statistical inference is conditional on the analysts’ allocation scheme. The method proposed here is a simple, yet far more defensible, alternative to these schemes: allow individuals to contribute as many observations to the sample as is necessary to exhaust their multiple identifications, effectively shifting the observational unit from a respondent to an identification, and then apply a post hoc correction to the violation of the *iid* assumptions. I show that a correction developed for complex survey designs can be appropriated for analyses involving multiple racial identifications. The paper concludes with an empirical demonstration of the consequences of arbitrary allocation schemes and an application of the correction method.

In the spring of 2000, the U.S. Census Bureau distributed census questionnaire forms on which respondents were asked to “**mark : one or more races** to indicate what this person considers himself/herself to be.” While this question seems innocuous enough, and it brings the Bureau’s data collection into compliance with revised standards for the classification of race and ethnicity issued by the Office of Management and Budget in October 1997, as Census 2000 data products are released over the next couple of years data analysts will be confronted with the prickly question of how to code multiple racial self-identifications. Some statistical packages, such as SPSS, have multiple response procedures that can tabulate a univariate distribution by treating each category as a binary response, e.g., *White* (yes or no); *Black or African American* (yes or no); *American Indian or Alaska Native* (yes or no); etc., but the utility of this tabulation is limited to descriptive reports based on the total number of responses in each category with the sum of all responses exceeding 100 percent of the census count. Alternatively, and more likely, census data analysts will make arbitrary decisions to allocate multi-racial respondents to single-race categories. Treated in this manner, race may be employed as a nominal-level covariate of any number of interesting and important social, economic, or demographic dependent variables, but statistical inference will be conditional on the analysts’ allocation scheme. Indeed, as will be shown in this paper, two analysts using the same data and the same model may reach two different conclusions because of their different allocation schemes.

When entered on the right-hand side of a prediction equation, multiple racial identifications pose a vexing challenge to individual-level analyses. Typically, analysts make arbitrary allocation decisions that either result in the loss of important information or corrupt the empirical distribution, or both. The method proposed here is a simple, yet far more defensible, alternative when multiple identifications are endemic to the analysis: allow individuals to

contribute as many observations to the sample as is necessary to exhaust their multiple identifications, effectively shifting the observational unit from an individual to an identification, and then apply a post hoc correction to the violation of the *iid* assumptions. In this paper, I will show that a correction developed for complex survey designs can be appropriated for analyses involving multiple racial identifications. The paper concludes with an empirical demonstration of the consequences of arbitrary allocation schemes and an application of the correction method.

Design Effects in Complex Surveys

Survey practitioners have long been in the business of designing economically efficient sampling strategies. Since the individual selection of elements is generally needlessly expensive, the sampling of units containing several elements can facilitate cost-efficient survey designs. These units are known as *clusters*. In Census 2000, by allowing respondents to mark one or more of six racial categories, respondents may create *natural clusters* (i.e., clusters that were not part of the survey design) ranging in size from two to six.

Generally, the homogeneity exhibited within clusters tends to increase the variance of sample estimators and estimates of variance (wrongly) based on *iid* (observations are independent and identically distributed) assumptions are downwards biased.¹ Consequently, the standard errors of descriptive statistics are too small and test statistics based on downwardly biased estimates of variance appear to be more significant than is really the case. These issues led Leslie Kish to develop the *design effect* as a means of quantifying this bias. Intuitively, the design effect may be understood as a ratio of the variance obtained under a complex survey design (e.g., cluster sampling) to the variance that would have been obtained if observations were collected through simple random sampling (Kish 1965).

Although Kish, and others, continued to revise and refine the methods for design effects after its conception in the late-1950's (cf. Kish 1957; Kish and Frankel 1974; Binder 1983; Skinner, Holt, and Smith 1989; Kish 1995), the means through which the design effect (*deff*) quantifies violation of the *iid* assumptions can probably best be seen in its earliest formulation (Kish 1965: 161-164):

$$deff = \frac{s_a^2/a}{s^2/n} = [1 + rho(b-1)].$$

The rightmost quantity, $[1 + rho(b-1)]$, where *rho* is an intra-cluster correlation coefficient that describes the degree of cluster homogeneity and *b* is the number (or average number) of observations per cluster, shows *deff* to be an estimate of the effect of clustering on sample design. As the ratio of a properly computed variance, s_a^2/a , to the variance under simple random sampling, s^2/n , of the same size, *deff* can be thought of as a multiplier, or correction factor, on variance calculations when a complex survey design is ignored. As the ratio of these two variance estimates is core to the definition of the design effect, it will be useful to expand on their derivations.

Begin by defining *N* as the number of elements in the population, *n* as the number of elements in the sample, and assuming an equal probability of selection method for the elements (EPSEM). If y_i is the value of the Y_j variable for the i^{th} sample element, then $y = \sum_i^n y_i$ gives the

simple sample total for the Y_j variable, $\bar{y} = \frac{y}{n} = \frac{1}{n} \sum_i^n y_i$ gives the simple sample mean, and

$s^2 = \frac{1}{n-1} \sum_i^n (y_i - \bar{y})^2$ gives the variance of the sample elements. In simple random samples,

\bar{y} and s^2 are unbiased estimators of \bar{Y} and S^2 , the population mean and variance, respectively (see Kish 1965: 35-36). We can further define the simple mean of the sample of an srs (simple random sample) selection, what Kish refers to as the *srs mean* and what is more generally referred to in introductory statistics texts as *mean of the sampling distribution*, as \bar{y}_0 , using the subscript 0 to distinguish it from the sample mean above, and the *variance of the srs mean* is computed as $\text{var}(\bar{y}_0) = \frac{s^2}{n}$ (Kish 1965: 40-41).²

Now suppose that from a population of A clusters, a sample clusters are selected with epsem (equal probability); and in the selected clusters, all B elements are included in the sample (which consists of $a \times B = n$ elements). The simple sample mean \bar{y} (from above) of the n elements in the sample still serves as an unbiased estimate the population mean \bar{Y} , but it is also the mean of the a cluster means:

$$\bar{y} = \frac{y}{n} = \frac{1}{n} \sum_j y_j = \frac{1}{aB} \sum_a \sum_b y_{ab} = \frac{1}{aB} \sum_a y_a = \frac{1}{a} \sum_a \bar{y}_a .$$

Assume further that the a clusters are selected with simple random choice from the list of A population clusters. For this design, the variance is stated as:

$$\text{var}(\bar{y}) = \frac{s_a^2}{a}, \quad \text{where } s_a^2 = \frac{1}{a-1} \sum_a (\bar{y}_a - \bar{y})^2 ,$$

again omitting the finite population correction (Kish 1965: 151-153). Whereas the mean \bar{y} may be computed simply from the entire sample, computing $\text{var}(\bar{y})$ requires the separation of the a cluster values. The importance of this point is hard to overstate. It means that if one ignores the complex survey design (e.g., a cluster design), descriptive (and test) statistics are unbiased (as point estimates), but estimates of variances of the sampling distribution are biased. In other

words, generally $\frac{s_a^2}{a} \neq \frac{s^2}{n}$ and $deff \neq 1$. While $deff$ quantifies the discrepancy between the variance estimates, $deff$ is generally more useful in its application as a correction factor. If

$\frac{s^2}{n}$ is the “conventional” estimate of $\text{var}(\bar{y})$, $\frac{s_a^2}{a}$ is the “properly computed” estimate of

$\text{var}(\bar{y})$, and $deff = \frac{\frac{s_a^2}{a}}{\frac{s^2}{n}} = [1 + rho(b-1)]$, then $\frac{s^2}{n} * [1 + rho(b-1)]$ makes the “corrected”

variance estimate. Clearly, as the degree of cluster homogeneity increases, that is, as rho goes to unity, $deff$ adjusts the variance of the srs mean upwards. In the absence of cluster homogeneity, i.e., rho is equal to zero, $deff$ is equal to one, and no adjustment is made.³ Also, all else equal, as cluster size increases, $deff$ increases.

Alternative Estimators of DEFF

In his seminal book on the statistical treatment of quantal assay data, Finney (1962 [1947]) observed that if a batch of n subjects is exposed to the dose I_0 , and all react independently, the probability of exactly r responding is the *Binomial Distribution* of probabilities, $\frac{n!}{r!(n-r)!} P^r Q^{n-r}$. However, the reactions of separate members of a batch to the stimulus of a particular dose are not always independent. A correlation of response may result from incomplete randomness of selection of the batch or from unsatisfactory control of experimental conditions causing the number responding to be seriously affected by some factor other than the dose. Whatever the cause, the variance of the numbers responding to a given dose will not be given by the binomial distribution. Finney referred to this correlation of response as

heterogeneity, and such heterogeneity must make the weight to be attached to the data other than is appropriate to the binomial distribution (Finney 1962:14-16).

When a probability model (e.g., the probit or logit model) is fitted to n binomial proportions is satisfactory, the residual deviance has an approximate χ^2 -distribution on $(n - p)$ degrees of freedom, where p is the number of unknown parameters in the fitted model. Since the expected value of a χ^2 random variable on $(n - p)$ degrees of freedom is $(n - p)$, it follows that the residual deviance for a well-fitting model should be approximately equal to its number of degrees of freedom (i.e., the mean deviance should be close to one. When the probability model is thought to be correct, but the residual mean deviance either exceeds or fails to reach unity, the assumption of binomial variability may not be valid and the data are said to exhibit either *overdispersion* or *underdispersion* (Collett 1999:188), or, as Finney remarked, *heterogeneity*. The question then is how to model overdispersed or underdispersed response probabilities?⁴

Suppose that the i^{th} of n sets of binary data consists of y_i successes in n_i observations. In other words, the data consist of n observed proportions, y_i/n_i , $i = 1, 2, \dots, n$. Let $R_{i1}, R_{i2}, \dots, R_{in_i}$, be the random variables associated with the n_i observations in each set, where $R_{ij} = 1$, for $j = 1, 2, \dots, n_i$, corresponds to a success, and $R_{ij} = 0$ to a failure. We define the probability of a success as p_i , so $P(R_{ij} = 1) = p_i$. Since R_{ij} is a Bernoulli random variable, it is well-known that $E(R_{ij}) = p_i$ and $Var(R_{ij}) = p_i(1 - p_i)$. The number of successes, y_i , is then the observed value of the random variable $\sum_{j=1}^{n_i} R_{ij}$, and so $E(y_i) = \sum_{j=1}^{n_i} E(R_{ij}) = n_i p_i$ and the variance of y_i is given by:

$$Var(y_i) = \sum_{j=1}^{n_i} Var(R_{ij}) + 2 \sum_{j=1}^{n_i} \sum_{k \neq j}^{n_i} Cov(R_{ij}, R_{ik})$$

where $Cov(R_{ij}, R_{ik})$ is the covariance between R_{ij} and R_{ik} , for $j \neq k$, and $k = 1, 2, \dots, n_i$. If the n_i random variables, $R_{i1}, R_{i2}, \dots, R_{in_i}$ are mutually independent, each of these covariance terms will be zero. Suppose, however, a nonzero (intracluster) correlation between R_{ij} and R_{ik} :

$$\rho = \frac{Cov(R_{ij}, R_{ik})}{\sqrt{Var(R_{ij})Var(R_{ik})}} \quad .$$

Since $Var(R_{ij}) = Var(R_{ik}) = p_i(1 - p_i)$, it follows that $Cov(R_{ij}, R_{ik}) = \rho * p_i(1 - p_i)$, and so

$$\begin{aligned} Var(y_i) &= \sum_{j=1}^{n_i} p_i(1 - p_i) + 2 \sum_{j=1}^{n_i} \sum_{k \neq j} \rho * p_i(1 - p_i) \\ &= n_i p_i(1 - p_i) + n_i(n_i - 1)[\rho * p_i(1 - p_i)] \\ &= n_i p_i(1 - p_i)[1 + \rho * (n_i - 1)] \quad . \end{aligned}$$

When there is no (intracluster) correlation between pairs of binary observations, ρ is equal to zero and $Var(y_i) = n_i p_i(1 - p_i)$, the variance of y_i under binomial sampling. It is also clear that when observations are correlated (i.e., $\rho \neq 0$), this expression for the “properly computed variance” of y_i , for binary data, is exactly the corrected variance estimator derived by Kish for continuous y_i , where n_i gives the cluster size.

In practice, if we assume each observed proportion, p_i , is based on the same number of binary observations, that is, the case of equal n_i (say, n_0), the expected value of the χ^2 -statistic can be approximated by $(n - p)[1 + \rho * (n_0 - 1)]$, where p is the number of unknown parameters in a model fitted to the n proportions. If the $E(X^2) \approx (n - p)[1 + \rho * (n_0 - 1)]$, this suggests that now X^2 has a $[1 + \rho * (n_0 - 1)]\chi_{n-p}^2$ -distribution, where χ_{n-p}^2 denotes a chi-squared random variable with $(n - p)$ degrees of freedom. As a consequence, $[1 + \rho * (n_0 - 1)]$ can be approximated by dividing the deviance (or Pearson chi-square) for a particular model by $(n - p)$,

its degrees of freedom. In some statistical packages, e.g., SAS, GLIM, this correction factor is called the *scale parameter*.

Although the discussion here of *alternative estimators of deff* was limited to logit and probit probability models, the scale parameter is also routinely applied to models for event counts, often referred to as the *modified poisson regression model* (cf. Beck and Tolnay 1995).

Design Effects for Statistics and Generalized Linear Models

As Kish and others accumulated empirical evidence, it became known that design effects vary greatly within a single survey. Some variables can have much larger values of *deff* so it is desirable to calculate *deff* for all variables of interest and, perhaps more importantly, for subpopulations. It also became evident that values of $deft(\bar{y})$ were most useful for generalizing

to $deft(\mathbf{b})$ for other statistics (\mathbf{b}), $deft(\mathbf{b}) = \sqrt{\frac{\text{var}(\mathbf{b})}{\text{SRS var}(\mathbf{b})}}$, where SRS var(\mathbf{b}) is the variance

estimate given by the conventional estimator and var (\mathbf{b}) is the proper variance (Kish 1995).⁵

Obviously a highly flexible, and easy to implement, method for calculating design effects for a wide range of variables is needed.

The unified theory of generalized linear models, and the iterative methods of estimation used to fit these models, makes possible the estimation of $deft(\mathbf{b})$ for a wide class of models that includes discrete and continuous response variables.

Generalized linear models (GLMs) are specified by three components: a random component (which identifies the probability distribution of the response variable), a systematic component (which specifies a linear function of the explanatory variables), and a link describing the functional relationship between the systematic component and the expected value of the random component. The same algorithm applies regardless of the choice of distribution for the

random component or the choice of link function (Agresti 1990:80-83; see also Liao 1994). To fix ideas, consider the regression model:

$$y_i = x_i \mathbf{b} + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \mathbf{s}^2),$$

where the i^{th} observation y_i is assumed to be a realization of a random variable Y_i whose expected values are given by $\mathbf{m} = E(Y_i)$, $i = 1, \dots, N$. In the linear regression model we specify the expectation of the random variable Y (dropping the i subscript) as a linear combination of K unknown parameters and explanatory variables ($K=1$ above):

$$E(Y) = \mathbf{\hat{\eta}} = \sum_{k=1}^K b_k x_k .$$

To create a more general model, we introduce the variable, $\boldsymbol{\zeta} = (\mathbf{h}_1, \dots, \mathbf{h}_N)'$, which links the function $\sum_{k=1}^K b_k x_k$ to $\mathbf{\hat{\eta}}$. In matrix notation, $\boldsymbol{\zeta} = \mathbf{X}\hat{\mathbf{a}}$, where the vector $\boldsymbol{\zeta}$ is called the *linear predictor* produced by $\mathbf{X} (=x_1, \dots, x_K)$. \mathbf{X} is an $N \times K$ model matrix of values of the explanatory variables for the N observations (when the explanatory variables are categorical, \mathbf{X} is referred to as a “design matrix” and consists of either dummy variable coding or effect coding), and \mathbf{b} is a $1 \times K$ vector of model parameters. Regardless of the type of model, the set of explanatory variables always linearly produce $\boldsymbol{\zeta}$, which is a predictor of Y . The function of the relation, $\boldsymbol{\zeta} = g(\mathbf{\hat{\eta}})$, between $\boldsymbol{\zeta}$ and $\mathbf{\hat{\eta}}$, however, is to be specified. There are many possible link functions between $\boldsymbol{\zeta}$ and $\mathbf{\hat{\eta}}$. The link between $\boldsymbol{\zeta}$ and $\mathbf{\hat{\eta}}$ distinguishes one member of the GLMs from another. In the regression model above, the function $g(\mathbf{\hat{\eta}}) = \mathbf{\hat{\eta}}$ gives the *identity link* $\boldsymbol{\zeta} = \mathbf{\hat{\eta}}$, specifying the classical linear model for the mean response. The choice of link function, or statistical model, depends on the distribution of the data and theory. Specifically, the distribution of the random component in Y determines the link function and the type of GLM. Most of the

commonly used prediction models in the social and behavioral sciences are members of this unified theory, including classical linear regression; probit, logit, and multinomial logit regression; poisson regression; and loglinear analysis of contingency tables.

Variance Estimators from Generalized Linear Models

Generalized linear models fit a wide class of models that includes discrete and continuous response variables using the same algorithm regardless of link function and choice of distribution for the random component. Except for the most simple linear/normal models, this fit is obtained through iteration, typically using the Newton-Raphson algorithm and Fisher-scoring or the method of iteratively re-weighted least squares.⁶ The likelihood maximization problem is to obtain (maximum likelihood) estimates of the parameter vector \mathbf{b} :

$$\max_{\mathbf{b}} L(\mathbf{b}; \mathbf{X}), \text{ or equivalently, } \max_{\mathbf{b}} \ln L(\mathbf{b}; \mathbf{X})$$

where \mathbf{X} is an $N \times K$ data matrix (as above). Typically we assume that observations are *iid* and write the likelihood as:

$$\max_{\mathbf{b}} \ln L(\mathbf{b}; \mathbf{X}) = \max_{\mathbf{b}} \ln \ell(\mathbf{b}; \mathbf{x}_1) + \ln \ell(\mathbf{b}; \mathbf{x}_2) + \dots + \ln \ell(\mathbf{b}; \mathbf{x}_N).$$

Suppose for the moment that we obtain \mathbf{b} satisfying:

$$\max_{\mathbf{b}} \ln \ell(\mathbf{b}; \mathbf{x}_1) + \ln \ell(\mathbf{b}; \mathbf{x}_2) + \dots + \ln \ell(\mathbf{b}; \mathbf{x}_N).$$

The estimated variance of \mathbf{b} is given by $-\mathbf{H}^{-1}$, where \mathbf{H} is the matrix of second derivatives (called the Hessian):

$$\mathbf{H} = \frac{\partial^2 \ln L(\mathbf{b}; \mathbf{X})}{\partial \mathbf{b} \partial \mathbf{b}'} = \frac{\partial^2 \ln \ell(\mathbf{b}; \mathbf{x}_1)}{\partial \mathbf{b} \partial \mathbf{b}'} + \dots + \frac{\partial^2 \ln \ell(\mathbf{b}; \mathbf{x}_N)}{\partial \mathbf{b} \partial \mathbf{b}'}$$

and the square root of the diagonal of $-\mathbf{H}^{-1}$ are the estimated standard errors of \mathbf{b} .

Before we see why this is so, it is important to note that we have not made anything of the distinction between $\hat{\mathbf{a}}$, the estimates we obtain, and \mathbf{a} , the true values, when referring to the vector parameter \mathbf{b} . What we really want to show is that our *estimator* of \mathbf{a} has variance $-\mathbf{H}^{-1}$. There are important consequences to pursuing this distinction. First, the results shown in this section are asymptotic, meaning that in finite samples, $\hat{\mathbf{a}}$ may be a biased estimate of \mathbf{a} . Second, the variance estimate $-\mathbf{H}^{-1}$ is guaranteed to be the variance of $\hat{\mathbf{a}}$ only when $\hat{\mathbf{a}}$ becomes very close to \mathbf{a} . For this to be so, we must assume the likelihood function $L(\mathbf{b};\mathbf{X})$ is the true likelihood of the data. This is rarely the case, due to misspecification of the covariates, the link function, or the probability distribution function. Fortunately, the very problem that led us into this discussion, i.e., complex survey designs and our creation of “natural” clusters by using identifications as our units of analysis rather than individuals, is going to lead us to a variance estimator that does not need $L(\mathbf{b};\mathbf{X})$ to be the true density function for \mathbf{X} . We will return to this point momentarily.

For now, we return to the problem of showing that the estimated variance of \mathbf{b} is given by

$-\mathbf{H}^{-1}$, let $D = \frac{\partial}{\partial \mathbf{b}}$ and $D^2 = \frac{\partial^2}{\partial \mathbf{b} \partial \mathbf{b}'}$, we can then redefine \mathbf{H} as:

$$\mathbf{H} = D^2 \ln L(\mathbf{b};\mathbf{X}) = D^2 \ln \ell(\mathbf{b};\mathbf{x}_1) + \dots + D^2 \ln \ell(\mathbf{b};\mathbf{x}_N).$$

Writing L for $L(\mathbf{b};\mathbf{X})$, if one is willing to assume $L(\cdot)$ is the true density function of \mathbf{X} , it can be proven that:

$$E(D \ln L) = 0 \text{ and } -E(D^2 \ln L) = E((D \ln L)^2) \quad ,$$

where $E(\cdot)$ denotes the expectation. $D \ln L$ is called the *score vector* or *gradient vector*. To show that this vector is a function of \mathbf{b} , it is often written as \mathbf{g} , where \mathbf{g} is:

$$\mathbf{g}(\mathbf{b}; \mathbf{X}) = D \ln L(\mathbf{b}; \mathbf{X}) = \frac{\partial \ln L(\mathbf{b}; \mathbf{X})}{\partial \mathbf{b}}$$

and the expectations from above are given as $E(\mathbf{g}) = 0$ and $-E(\mathbf{H}) = E(\mathbf{g}\mathbf{g}')$. In other words, the score function has a mean of 0, since $E(\mathbf{g}) = 0$, and variance given by

$Var(\mathbf{g}) = E(\mathbf{g}\mathbf{g}') - E(\mathbf{g})E(\mathbf{g})' = E(\mathbf{g}\mathbf{g}') = -E(\mathbf{H})$, again because $E(\mathbf{g}) = 0$. Because \mathbf{g} is a function of \mathbf{b} , we can also obtain the variance of \mathbf{g} from:

$$Var(\mathbf{g}) \approx (D\mathbf{g})Var(\mathbf{b})(D\mathbf{g})' = (D^2 \ln L)Var(\mathbf{b})(D^2 \ln L)',$$

then, substituting from above, $\mathbf{H} = D^2 \ln L$, we can show $Var(\mathbf{g}) \approx \mathbf{H}Var(\mathbf{b})\mathbf{H}$ and rearranging yields $Var(\mathbf{b}) \approx \mathbf{H}^{-1}Var(\mathbf{g})\mathbf{H}^{-1}$. Finally, if indeed the $E(\mathbf{H})$ is approximately \mathbf{H} at the observed \mathbf{X} , we can substitute $-\mathbf{H}$ for $Var(\mathbf{g}) = -E(\mathbf{H})$ and have:

$$Var(\mathbf{b}) \approx \mathbf{H}^{-1}(-\mathbf{H})\mathbf{H}^{-1} = -\mathbf{H}^{-1},$$

which is what was claimed above.

Now, returning to our maximization problem, given the problem, $\max_{\mathbf{b}} \ln L(\mathbf{b}; \mathbf{X})$, how do we obtain the solution? The analytic solution for a simple problem would involve taking derivatives and setting them to zero, $\frac{\partial \ln L(\cdot)}{\partial \mathbf{b}} = 0$. In general, however, $\frac{\partial \ln L(\cdot)}{\partial \mathbf{b}} = 0$ is too complicated to admit an analytic solution and so we convert the maximization problem into a numerical maximization problem. Put simply, to find \mathbf{b} such that $f(\mathbf{b}) = \ln L(\mathbf{b}; \mathbf{X})$ is maximized,

1. Start with a guess \mathbf{b}_0 (called an *initial value*).
2. Calculate a direction vector $\mathbf{d} = \mathbf{g}(-\mathbf{H})^{-1}$, where \mathbf{g} is the aforementioned *score vector* and $-\mathbf{H}^{-1}$ is variance estimate of \mathbf{b} .
3. Calculate a new guess $\mathbf{b}_1 = \mathbf{b}_0 + s\mathbf{d}$, where s is a scalar.
 - a. Start with $s = 1$.

- b. If $f(\mathbf{b}_0 + \mathbf{d}) > f(\mathbf{b}_0)$, try $s = 2$. If $f(\mathbf{b}_0 + 2\mathbf{d}) > f(\mathbf{b}_0 + \mathbf{d})$, try $s = 3$, and so on.
 - c. If $f(\mathbf{b}_0 + \mathbf{d}) \leq f(\mathbf{b}_0)$, back up and try $s = .5$, etc.
4. Repeat.

In other words, the maximizing routine needs to specify the function, $f(\mathbf{b}) = \ln L(\mathbf{b}; \mathbf{X})$, its first

derivatives, $\mathbf{g}(\mathbf{b}) = \frac{\partial f}{\partial \mathbf{b}}$, and its second derivatives, $\mathbf{H}(\mathbf{b}) = \frac{\partial^2 f}{\partial \mathbf{b} \partial \mathbf{b}'}$.

The result that $Var(\hat{\mathbf{a}})$ is asymptotically $-\mathbf{H}^{-1}$ is for the true value $\hat{\mathbf{a}}$ of \mathbf{b} only, but what if $L(\mathbf{b}; \mathbf{X})$ is not the true likelihood for \mathbf{X} ? Indeed, we know this will be so because our analysis is based on identifications rather than individuals, while likelihood theory assumes that observations are *iid*. Luckily, we can derive an empirical variance estimator, alternatively called the *robust*, *Huber-White*, or *sandwich*, variance estimator (cf. Huber 1967; White, 1980; 1982; Binder 1983) for $\mathbf{b} = \hat{\mathbf{a}}$.

Recall that we showed $Var(\mathbf{b}) \approx \mathbf{H}^{-1} Var(\mathbf{g}) \mathbf{H}^{-1}$. We want to evaluate the formula at $\mathbf{b} = \hat{\mathbf{a}}$, so we will write $Var(\hat{\mathbf{a}}) \approx \mathbf{H}(\hat{\mathbf{a}})^{-1} Var(\mathbf{g}(\hat{\mathbf{a}})) \mathbf{H}(\hat{\mathbf{a}})^{-1}$. We also showed the score vector, $\mathbf{g}(\mathbf{b}; \mathbf{X})$, now written as $\mathbf{g}(\hat{\mathbf{a}}; \mathbf{X})$,

$$= \frac{\partial \ln L(\hat{\mathbf{a}}; \mathbf{X})}{\partial \mathbf{b}} = \frac{\partial \ln \ell(\hat{\mathbf{a}}; \mathbf{x}_1)}{\partial \mathbf{b}} + \dots + \frac{\partial \ln \ell(\hat{\mathbf{a}}; \mathbf{x}_N)}{\partial \mathbf{b}}.$$

This shows the score vector, \mathbf{g} , is just the sum of N *iid* random variables.

Any intermediate statistics text will give the variance of a sum as simply:

$$Ns^2 = \frac{N}{N-1} \sum_{i=1}^N (z_i - \bar{z})^2, \quad ,$$

where we use $z_i = \mathbf{g}(\hat{\mathbf{a}}; \mathbf{x}_i)$ to obtain an estimate of $Var(\mathbf{g}(\hat{\mathbf{a}}))$. Since $E(\mathbf{g}(\hat{\mathbf{a}}; \mathbf{X})) = 0$, as shown earlier, we can substitute $\bar{e} = 0$ and plugging the empirical estimate of $Var(\mathbf{g}(\hat{\mathbf{a}}))$ into

$Var(\hat{\mathbf{a}}) \approx \mathbf{H}(\hat{\mathbf{a}})^{-1} Var(\mathbf{g}(\hat{\mathbf{a}})) \mathbf{H}(\hat{\mathbf{a}})^{-1}$ gives:

$$Var(\hat{\mathbf{a}}) \approx \mathbf{H}(\hat{\mathbf{a}})^{-1} \left(\frac{N}{N-1} \sum_{i=1}^N \mathbf{g}_i \mathbf{g}_i' \right) \mathbf{H}(\hat{\mathbf{a}})^{-1} .$$

In all generalized linear models, it is only the linear form $\mathbf{x}_i \mathbf{b}$ that enters the function L_i . In these cases,

$$\frac{\partial \ln L_i(\mathbf{x}_i, \mathbf{b})}{\partial \mathbf{b}} = \frac{\partial \ln L_i(\mathbf{x}_i, \mathbf{b})}{\partial (\mathbf{x}_i \mathbf{b})} \frac{\partial (\mathbf{x}_i \mathbf{b})}{\partial \mathbf{b}} = \frac{\partial \ln L_i(\mathbf{x}_i, \mathbf{b})}{\partial (\mathbf{x}_i \mathbf{b})} \mathbf{x}_i ,$$

so, writing $\mathbf{g}_i = \frac{\partial \ln L_i(\mathbf{x}_i, \mathbf{b})}{\partial (\mathbf{x}_i \mathbf{b})}$ becomes simply $\mathbf{g}_i \mathbf{x}_i$. Thus, the formula for the robust estimate of

variance can be rewritten:

$$\hat{\mathbf{O}} = \hat{\mathbf{V}} \left(\sum_{i=1}^N g_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \hat{\mathbf{V}}$$

where $\hat{\mathbf{V}} = -\mathbf{H}^{-1} = -\left(\frac{\partial^2 \ln L}{\partial \mathbf{b} \partial \mathbf{b}'} \right)^{-1}$ is the ‘‘conventional’’ estimator of variance and \mathbf{g}_i (a row vector)

is the contribution from the i^{th} observation to the scores $\frac{\partial \ln L}{\partial \mathbf{b}}$. When referred to in the

singular, g_i , is called the *score*, with $\sum_i g_i = 0$ and $COV(\mathbf{g}_i, \mathbf{x}_i) = 0$, calculated over $i=1, \dots, N$.

Indeed, in all applications of GLMs, e.g., classical linear regression, logit regression,

multinomial logit regression, g_i is simply the residual (the difference between the observed and predicted value of y_i) for the i^{th} observation and the usual expression (cf. White 1980) becomes:

$$\hat{O} = \hat{V} \left(\sum_{i=1}^N e_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \hat{V} \quad .$$

The inner-product of the above estimator, $\sum_{i=1}^N e_i^2 \mathbf{x}_i \mathbf{x}_i'$, is still only appropriate when observations are independent. For clustered data (and complex survey data), this estimator must be replaced by one appropriate for the independent units of the data. Since clusters (or PSUs in complex survey designs) are independent, we can sum the scores within a cluster to create a *super-observation* and then use the formula for the variance of a sum on these independent super-observations. Hence the *cluster robust variance estimator* becomes:

$$\hat{O} = \hat{V} \left[\sum_{m=1}^{n_c} \left(\sum_{i \in C_m} e_i \mathbf{x}_i \right) \left(\sum_{i \in C_m} e_i \mathbf{x}_i \right)' \right] \hat{V} \quad ,$$

where C_m contains the indices of the observations belonging to the m^{th} cluster for $m = 1, 2, \dots, n_c$ with n_c total clusters.⁷ The *cluster robust variance estimator* is thus an appropriate substitution for the “properly computed variance” sought for the numerator of *deff*.

“A Properly Computed Variance”

The robust variance estimator may also be derived directly from properties of the linear regression model. Indeed this approach is instructive because it permits a more intuitive interpretation than is readily apparent above. Written in terms of individual observations, the variance of the slope in a simple regression model is given by:

$$\text{VAR}(\hat{\mathbf{b}}) = E \left(k_1^2 e_1^2 + k_2^2 e_2^2 + \dots + k_n^2 e_n^2 + 2k_1 k_2 e_1 e_2 + \dots + 2k_{n-1} k_n e_{n-1} e_n \right) \quad ,$$

where $k_i = \frac{x_i}{\sum x_i^2}$ and $e_i = y_i - \hat{y}_i$, $i = 1, \dots, n$. Since the expectations of the cross-product terms

are zero (because of the assumption of no serial correlation) and the k_i are known constants:

$$\text{VAR}(\hat{\mathbf{b}}) = k_1^2 E(e_1^2) + k_2^2 E(e_2^2) + \dots + k_n^2 E(e_n^2) \quad .$$

Under the assumption of constant variance, $E(e_i^2) = \mathbf{s}^2$, and the above definition of k_i , we get

the familiar expression for the variance of a slope:

$$\text{VAR}(\hat{\mathbf{b}}) = \sum k_i^2 \mathbf{s}^2 = \frac{\mathbf{s}^2 \sum x_i^2}{(\sum x_i^2)^2} = \frac{\mathbf{s}^2}{\sum x_i^2} \quad .$$

If we relax the constant variance assumption, $E(e_i^2) = \mathbf{s}_i^2$, $\text{VAR}(\hat{\mathbf{b}})$ becomes:

$$\frac{\sum x_i^2 \mathbf{s}_i^2}{(\sum x_i^2)^2} \quad ,$$

precisely the expression of the robust variance estimator given in matrix notation in the previous section. Although the cluster robust variance estimator accomplishes what is intended of Kish's *deft(b)*, that is, to correct for sample clustering, *deft(b)* itself may be of interest in that it provides a convenient quantification of the effect of clustering (or sample design) on the precision of the parameter vector. This quantity is given by the square root of the ratio of the diagonal elements of $\hat{\Sigma}$, the cluster robust variance estimates, to the corresponding elements from $\hat{\mathbf{V}}$, the conventional variance estimates.

It is also worth highlighting the fact that the form of the *cluster robust variance estimator* ensures that any amount of additional clustering within the primary cluster, or, in the case of complex survey designs, secondary clustering within the primary sampling units will have no effect on variance estimates. This issue becomes relevant in analyses with multiple identifications from complex sample surveys with PSUs (in which the respondents generating

multiple responses are embedded). In these situations, the “natural” clusters created by allowing individuals to contribute multiple observations are actually of no consequence to the *cluster robust variance estimates*. To see this, we only have to understand that any degree of summing of \mathbf{g}_i within each of the n_c clusters will be lost on the final summation. A simple example will show this. Suppose we have two sampling designs, in the first we have two PSUs and no additional clustering. In the second design, we also have two PSUs but within each of these we have two secondary clusters (2SUs). For the sake of simplicity, assume that $g_i^2 = 1$ for each i observation:

		Design 1	Design 2
PSU id	2SU id	g_i^2	
1	1	1	1
1	1	1	1
	$\sum_{i \in 2SU_m} g_i^2$		2
1	2	1	1
1	2	1	1
	$\sum_{i \in 2SU_m} g_i^2$		2
$\sum_{j \in PSU_m} g_i^2$		4	4
2	3	1	1
2	3	1	1
	$\sum_{i \in 2SU_m} g_i^2$		2
2	4	1	1
2	4	1	1
	$\sum_{i \in 2SU_m} g_i^2$		2
$\sum_{j \in PSU_m} g_i^2$		4	4

Clearly, regardless of whether we first sum (squared) scores within the second-stage sampling units, $\sum_{i \in 2SU_m} g_i^2$, and then sum these summed 2SU (squared) scores within the primary sampling units or simply sum the (squared) scores within the PSUs, the totals, $\sum_{j \in PSU_m} g_j^2$, are identical.

Race as a Predictor of Annual Personal Income

Following the release of the Census 2000 microdata sample files, currently scheduled for 2002 and 2003, data analysts will be able to assess racial group differences among a host of outcome variables, adjusting for a variety of characteristics not permitted by earlier tabular releases and at geographical locations not revealed by smaller sample sets. Without a viable alternative, these analysts will make arbitrary decisions to allocate multi-racial respondents to single-racial categories. Using simulated data, I demonstrate the consequences of several allocation schemes and contrast these to the alternative method developed in this paper by regressing annual personal income on race, Hispanic origin, sex, age, education, and English language spoken at home.⁸ In this demonstration, individuals can contribute up to four racial self-identifications (*white, black, Asian, and some other race*). *Hispanic origin* is entered as a distinct identification of ethnicity. Treated in this manner, 4,870 individuals contribute a total of 5,296 identifications. From the perspective of individual respondents, 4,469 report only one race, 376 contribute two racial identifications, and 25 individuals self-identify with three racial groups. In other words, just over 8.2 percent of the 4,870 individuals self-identify as multi-racial (= $[401/4870] * 100$). Table 1 reports descriptive statistics based on these 5,296 identifications.

[Table 1 about here]

In the absence of an allocation scheme, it is only under a reporting of identifications that the proportions of races will sum to unity. On the other hand, the distribution of identifications results in duplications of all other characteristics for the 401 individuals who self-identify as multi-racial.

Four allocation schemes are investigated and the resultant racial distributions are given in Table 2. Under the first scheme, any individual self-identified as *white* will be allocated to that racial group, regardless of any additional identifications. Any individual self-identified as *black*, given that he/she did not also identify as *white*, will be allocated as *black*, regardless of any additional identifications as *Asian* or *some other race (SOR)*. Finally, any individual self-identified as *Asian*, given that he/she did not also identify as *white* or *black*, will be allocated as *Asian*, regardless of an additional *SOR* identification. Under this scheme, *SOR* allocation must be a single-racial identification. The second allocation scheme repeats this exercise, but makes *white* the lowest priority identification (and *black*, the highest). Under this scheme, *white* allocation must be a single-racial identification. The third scheme is derived from the suggested recoding of races by the National Longitudinal Study of Adolescent Health (Bearman, Jones, and Udry 1997) in which Hispanic origin is treated as a racial classification rather than as a distinct identification of ethnicity. Under this scheme, any individual self-identified as *Hispanic* will be assigned to this “race” group, regardless of any additional identifications. All self-identified *Non-Hispanics*, will be allocated according to the second scheme described above.⁹ The fourth scheme uses a random selection of one identity for each of the 401 multi-racial individuals.

[Table 2 about here]

As expected, the random allocation of a single identification for multi-racial individuals (allocation scheme 3) most closely resembles the allocation of identifications from Table 1

(reproduced in the final column of Table 2). And given the relatively small percentage of multi-racial individuals, it is not surprising that allocation schemes 1 and 2 also do not depart too severely from the total identification distribution. Allocation scheme 3, on the other hand, draws disproportionately from the *SOR* group, revealing the preference of many of *Hispanic origin* to designate themselves as a unique racial group, and (to a somewhat lesser degree) from the *white* racial group.

Table 3 shows the regression parameter estimates obtained using multiple identifications (with and without the cluster robust standard errors) and under each of the allocation schemes described above.¹⁰

[Table 3 about here]

While the magnitudes (and signs) of the effects are generally similar across estimation methods, of principal interest is the shaded area of Table 3. By allowing individuals to contribute multiple observations (via identifications) and then correcting the *iid* assumptions violation rather than applying an arbitrary allocation scheme, we would infer opposite conclusions concerning the statistical significance of the *Asian-white* and *SOR-white* group differences. Under any of the allocation schemes, we would infer a statistically significant difference in the adjusted incomes of *whites* and *Asians* and no difference in the adjusted incomes of *whites* and the *SOR* group. Using individuals' multiple identifications, and the cluster robust standard errors, we fail to reject the null of no difference between the adjusted incomes of *Asians* and *whites*, but we would infer a statistically significant difference between the *SOR* group and *whites*. In addition, a comparison of the corrected and uncorrected multiple identification standard errors (defined above as *deft(b)*) substantiates the desirability of statistic-specific corrections rather than a global dispersion scale parameter: in some instances the standard errors are adjusted upwards, in others

they are adjusted downwards; in some the adjustments are relatively small, and in others the adjustment is of great consequence.

To facilitate this observation, Table 4 reproduces the regression estimates of the corrected multiple identification column of Table 3 along with their design effects, $deff(b)$, and the *effective n* on which each inferential test is based, where *effective n* is defined as: $deff_n = n / deff(b)$.

[Table 4 about here]

Defined as multipliers on the precision (variance) of each estimated regression coefficient, $deff(b)$ greater than unity imply that the corrected standard errors are adjusted upwards, while $deff(b)$ less than one imply that the corrected standard errors are adjusted downwards. In three instances (*SOR*, *High School*, and *English*), the corrected standard errors are over three times smaller than the unadjusted standard errors. In terms of the *effective n*, $deff(b)$ in excess of one imply that the corrected standard errors are equivalent to those that would be given in srs selection with more observations than our total identifications and $deff(b)$ less than one imply that the corrected standard errors would be given by a smaller srs selection. Several of the $deff_n$ are larger than either the number of individuals or their identifications, while in other instances inference is based on $deff_n$ substantially smaller than the even the number of individuals.

Discussion

The regression analyses given in Table 3 show that, even with a relatively small proportion of multi-racial respondents, arbitrary allocation schemes for multiple racial identifications will affect statistical inference. The method presented in this paper provides an

alternative in which respondents determine their identities, and the resultant inference from these identities, rather than data analysts.

Following the 1990 U.S. Census, there has been a growing (rhetorical, if not constituent) demand that multi-racial identification is itself a classification that deserves attention. In the years leading up to Census 2000 there was a concerted effort to have a multi-racial category included among the principal races on the census form and indeed the Census Bureau pre-tested questionnaires of this type. Ultimately the Census Bureau did not include a multi-racial category on the census questionnaire, but the “**mark : one or more races**” option is clearly a significant move toward acknowledging multi-racial identity. Each of the allocation schemes described earlier could be modified to include a single multi-racial category, but this strategy would draw respondents out of the major racial categories, while obscuring the potentially unique behaviors of particular racial combinations. Alternatively, the analyst could create dummy variables for single races and for every multi-racial combination. This would require the analyst to include 62 dummy variables to capture all 63 combinations of the six major racial categories identified on the Census 2000 questionnaire forms. Clearly, this would be unwieldy and, even with large data sets, carry a high potential for creating identification problems.¹¹

It is easy to incorporate a multi-racial identification net of the single race dummy variables under the method advocated here. Indeed, if there is unique covariation between multi-racial status and the dependent variable, it will be reflected in the statistical significance of the t-ratio on this additional regressor. The final column in Table 4 shows the regression estimates of the multiple identification model adjusted by an additional dummy variable for multi-racial identity. In this application, inclusion of multi-racial identity neither improves the overall fit of

the model (i.e., the t-ratio is not statistically significant) nor does it alter in any substantive manner the effects of the other covariates.

Conclusion

Multiple racial identifications pose a vexing challenge to individual-level analyses. Typically data analysts apply allocation schemes that lose important information or corrupt the empirical distribution, or both. The problem, and this course of action, is of course more generic than analyses involving racial self-identifications. In this paper I show that a post hoc correction developed for complex survey designs can be appropriated for analyses involving multiple responses, thus freeing analysts from imposing arbitrary structures on their data.

Notes

¹ Although negative intra-cluster correlation is possible (this occurs when clusters are more uniform than would be produced by random sorting) and if present would result in the upwards bias of variance estimates.

² Note that we have omitted the *finite population correction* (fpc) factor, $(1 - f) = \left(1 - \frac{n}{N}\right)$, a correction factor for sampling without replacement, which is generally negligible (i.e., $f \sim 0$).

³ Consistent with our previous note, a negative intra-class correlation results in a downwards adjustment of variance estimates. The lowest possible value for ρ is $-1/(b-1)$. This corresponds to $[1 + \rho(b-1)] = 0$ and to zero variance between cluster means (Kish 1965: 163).

⁴ The following discussion in this section is drawn largely from Collett (1999), see esp. Pp. 192-200.

⁵ Up to this point in our discussion we have been referring only to the design effect as a multiplier on the variance of an srs mean. Kish and others recognized very early that the most common application would to be found in its generalization to test statistics. As such, a direct multiplier on the test statistic, or its standard error, is most useful. Kish defined this multiplier, $deft$, where $deft = \sqrt{deff}$ when $f \sim 0$ (Kish 1995).

⁶ Statistical software, such as SAS (SAS Institute Inc. 1999), GLIM (Francis, Green, and Payne 1994), and STATA (StataCorp 2001) use such methods to find the maximum likelihood estimates. The discussion in this section closely follows from Gould and Sribney (1999: 2-14) and, except for the material on the *robust variance estimator*, this material represents a standard introduction to maximum likelihood estimation and can be found in a number of texts on that

subject and in relevant sections in introductory calculus texts (cf. Francis, Green, and Payne 1994: 263-267; Eliason 1993; Iversen 1996).

⁷ Discussion of the *cluster robust variance estimator* can be found in the Stata Reference Manual Volume 3, see esp. Pp. 241-243 (StataCorp 1999).

⁸ Dr. Jorge H. del Pinal, Assistant Chief for Special Population Statistics at the U.S. Census Bureau, provided this data. As noted above, the data is simulated and is generated only for the demonstration of the method for multiple racial identifications described in this paper. It is not a reproduction of the social, economic, and demographic characteristics of any actual population.

⁹ This coding scheme is described at the Adolescent Health Project web site

[URL:<http://www.cpc.unc.edu/projects/addhealth/race.html>](http://www.cpc.unc.edu/projects/addhealth/race.html).

¹⁰ Program code for fitting the multiple identification models in Table 3 is given in Appendix 1. The first block of code fits the uncorrected multiple identification model using SAS' standard regression procedure (PROC REG). The estimated standard errors are then adjusted following the method described in this paper using SAS' IML procedure. Appendix 1 also implements the standard error correction using GEE estimation in SAS' GLM procedure (PROC GENMOD). Appendix 2 gives SAS code (using the GENMOD and IML procedures) to implement the standard error correction for a binomial logistic regression (of the dependent variable COLLEGE). Corrected standard errors are also easily obtained through STATA's SVYREG and SVYLOGIT procedures using the *robust* and *cluster* options.

¹¹ Even with only four racial categories, one would have to include 14 dummy variables to capture the 15 combinations in the empirical application here.

References

- Agresti, Alan. 1990. *Categorical Data Analysis*. New York: Wiley.
- Bearman, P. S., J. Jones, and J. R. Udry. 1997. *The National Longitudinal Study of Adolescent Health: Research Design* [WWW document].
- [URL:http://www.cpc.unc.edu/projects/addhealth/design.html](http://www.cpc.unc.edu/projects/addhealth/design.html).
- Beck, E. M., and Stewart E. Tolnay. 1995. "Analyzing Historical Count Data: Poisson and Negative Binomial Regression Models." *Historical Methods* 28: 125-131.
- Binder, David A. 1983. "On Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51: 279-292.
- Collett, D. 1999. *Modelling Binary Data*. Boca Raton: Chapman & Hall/CRC.
- Eliason, Scott. 1993. *Maximum Likelihood Estimation*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-96. Thousand Oaks, CA: Sage.
- Finney, D. J. 1962. *Probit Analysis*. Cambridge: Cambridge University Press.
- Francis, Brian, Mick Green, and Clive Payne. 1994. *The GLIM System: Release 4 Manual*. Oxford: Clarendon Press.
- Gould, William, and William Scribney. 1999. *Maximum Likelihood Estimation With Stata*. **College Station, TX: Stata Corporation.**
- Huber, P. J. 1967. "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions." Pp. 221-233 in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*. Berkeley, CA: University of California Press.
- Iversen, Gudmund R. 1996. *Calculus*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-110. Thousand Oaks, CA: Sage.
- Liao, Tim Futing. 1994. *Interpreting Probability Models: Logit, Probit, and Other Generalized*

-
- Linear Models*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-101. Thousand Oaks, CA: Sage.
- Kish, Leslie. 1957. "Confidence Intervals for Clustered Samples." *American Sociological Review* 22: 154-165.
- Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.
- Kish, Leslie. 1995. "Methods for Design Effects." *Journal of Official Statistics* 11:55-77.
- Kish, Leslie and Martin Richard Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society Series B (Methodological)* 36: 1-22.
- SAS Institute Inc. 1999. *SAS/STAT User's Guide, Version 8*. Cary, NC: SAS Institute Inc.
- Skinner, C. J., D. Holt, and T. M. F. Smith. 1989. *Analysis of Complex Surveys*. New York: Wiley.
- StataCorp. 1999. *Stata Statistical Software: Release 6.0*. College Station, TX: Stata Corporation.
- StataCorp. 2001. *Stata Statistical Software: Release 7.0*. College Station, TX: Stata Corporation.
- White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48: 817-830.
- White, H. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50: 1-25.

	Mean	Standard Deviation
Dependent Variable		
Annual Personal Income (in \$)	36586.01	47044.51
Independent Variables		
White	.43	
Black	.17	
Asian	.22	
Some Other Race	.18	
Hispanic	.32	
Sex (Female)	.45	
Age	29.63	3.78
High School Degree (+)	.69	
College Degree (+)	.38	
English Spoken in Home	.82	

	Allocation Scheme 1	Allocation Scheme 2	Allocation Scheme 3	Allocation Scheme 4	Allocation From Table 1
Racial Group	Percent	Percent	Percent	Percent	Percent
White	46.98	39.82	30.60	43.14	43.20
Black	15.93	18.52	16.10	16.92	17.03
Asian	20.88	22.85	22.05	22.11	21.62
Some Other Race	16.20	18.81	0.76	17.82	18.15
Hispanic			30.49		
N	4870	4870	4870	4870	5296

Table 3. Regression of Annual Personal Income on Race and Other Predictors Using Multiple Identifications and Under Allocation Schemes (standard errors given in parentheses)

Predictor Variables¹	Corrected Multiple Identifications	Uncorrected Multiple Identifications	Allocation Scheme 1	Allocation Scheme 2	Allocation Scheme 3	Allocation Scheme 4
Black	-8345.97 (1450.62)*	-8345.97 (1775.33)*	-9057.68 (1848.58)*	-8809.11 (1782.13)*	-9671.74 (1939.98)*	-9374.90 (1824.62)*
Asian	-3223.14 (1773.03)	-3223.14 (1697.53)	-4143.96 (1756.17)*	-4356.01 (1722.64)*	-4895.84 (1795.91)*	-3616.02 (1730.22)*
SOR	-3134.33 (1094.21)*	-3134.33 (2080.29)	-3492.70 (2196.12)	-3806.49 (2223.96)	-10243.00 (7217.01)	-4051.95 (2201.73)
Hispanic	-6012.56 (1387.64)*	-6012.56 (1861.83)*	-6254.25 (1944.07)*	-5800.31 (2027.33)*	-9468.31 (1779.51)*	-5696.66 (2000.65)*
Sex	-11844.00 (1465.62)*	-11844.00 (1236.82)*	-12035.00 (1266.39)*	-11919.00 (1267.22)*	-11959.00 (1265.55)*	-11964.00 (1266.72)*
Age	900.29 (194.28)*	900.29 (161.42)*	958.74 (164.90)*	952.50 (164.86)*	945.63 (164.85)*	961.02 (164.84)*
High School	6529.94 (891.62)*	6529.94 (1644.09)*	6959.23 (1687.23)*	7138.10 (1687.46)*	7034.42 (1686.86)*	7025.65 (1686.46)*
College	21778.00 (1850.99)*	21778.00 (1513.95)*	21436.00 (1555.67)*	21346.00 (1556.47)*	21303.00 (1556.80)*	21279.00 (1556.18)*
English	8007.14 (976.30)*	8007.14 (1789.78)*	7403.67 (1827.40)*	7622.63 (1819.16)*	7250.47 (1825.25)*	7740.56 (1820.53)*
Intercept	472.55 (5674.01)	472.55 (5306.68)	-744.49 (5422.12)	-548.03 (5421.66)	554.31 (5454.57)	-1039.97 (5418.22)
N	5296	5296	4870	4870	4870	4870

¹All predictor variables are dummy variables except age. Reference categories are: white (for black, Asian, and SOR), non-Hispanic, Male, less than high school degree (for high school and college), and English not spoken at home. Age is measured in years.

* t-ratio for two-tailed test statistically significant at $p < .05$

Table 4. Regression of Annual Personal Income on Race and Other Predictors Using Multiple Identifications (from Table 3).

Predictor Variables	Corrected Multiple Identifications	deff(b)	effective n (deff_n)	Multiple Identifications Plus Multi-racial Dummy
Black	-8345.97 (1450.62)*	0.67	3548.32	-8411.95 (1386.90)*
Asian	-3223.14 (1773.03)	1.09	5772.64	-3271.32 (1762.92)
SOR	-3134.33 (1094.21)*	0.28	1482.88	-3002.41 (1160.17)*
Hispanic	-6012.56 (1387.64)*	0.55	2912.80	-6330.23 (1590.80)*
Sex	-11844.00 (1465.62)*	1.40	7414.40	-11906.00 (1536.11)*
Age	900.29 (194.28)*	1.45	7679.20	900.89 (193.74)*
High School	6529.94 (891.62)*	0.29	1535.84	6461.45 (910.69)*
College	21778.00 (1850.99)*	1.49	7891.04	21809.35 (1884.51)*
English	8007.14 (976.30)*	0.30	1588.8	7869.65 (984.23)*
Multi-racial ¹				1435.29 (3095.93)
Intercept	472.55 (5674.01)	1.14	6037.44	506.27 (5703.24)
N	5296			5296

¹ Reference category is Single-Race Identification.
* t-ratio for two-tailed test statistically significant at $p < .05$

Appendix 1

*/

Use proc reg to fit income regression. Cluster robust standard errors are then calculated using output from proc reg by forming scores, summing within clusters (in proc summary), and reproducing the information matrix, its inverse, and the estimated covariance matrix of the parameter estimator in proc iml.

/*;

```
proc reg data=xxx;
  model income=age sex black asian sor hisp english hs coll;
  output out=score p=yhat r=ei;

data gi; set score;

int=1;

array score {10} gi1-gi10;
array xs {10} int age sex black asian sor hisp english hs coll;
do i=1 to 10;
  score{i}=xs{i} * ei;
end;

proc summary data=gi nway ;
  class id;
  var gi1 - gi10;
  output out=g sum= ;

data g; set g;

worksize=5000;
proc iml;
  use gi;
  read all var {int age sex black asian sor hisp english hs coll} into x;
  read all var {income} into y;
  use g;
  read all var {gi1 gi2 gi3 gi4 gi5 gi6 gi7 gi8 gi9 gi10} into gi;

  h=t(x)*x;
  xpxi=inv(t(x)*x);
  rv=xpxi*(t(gi)*gi)*xpxi;
  crse=sqrt(vecdiag(rv));

title 'cluster robust standard errors';
print crse;
```

*/

Use proc genmod (with repeated option) to fit income. GEE estimation in proc genmod through the repeated option gives cluster robust standard errors.

/*;

```
proc genmod data=xxx;  
class id;  
model income=age sex black asian sor hisp english hs coll;  
repeated subject=id;
```

Appendix 2

*/

Use proc genmod to fit binomial logistic regression of college education. The code in this program is very similar to that in appendix 1 except for the use of the obstats option to get values of the hessian weight and the incorporation of this weight into the information matrix in iml.

/*;

```
ods listing exclude obstats;
proc genmod data=xxx descending;
class id;
model coll=age sex black asian sor hisp english/dist=binomial obstats;
```

```
data gi;
drop pred xbeta std hesswgt upper lower resraw reschi resdev;
set score;
```

```
int=1;
w=hesswgt;
ei=resraw;
```

```
array score {8} gi1-gi8;
array xs {8} int age sex black asian sor hisp english;
do i=1 to 8;
  score{i}=xs{i} * ei;
end;
```

```
proc summary data=gi nway ;
class id;
var gi1 - gi8;
output out=g sum= ;
```

```
data g; set g;
```

```
worksize=5000;
proc iml;
use gi;
read all var {int age sex black asian sor hisp english} into x;
read all var {w} into w;
use g;
read all var {gi1 gi2 gi3 gi4 gi5 gi6 gi7 gi8} into gi;
```

```
w=diag(w);
h=t(x)*w*x;
xpxi=inv(h);
```

```
rv=xpxi*(t(gi)*gi)*xpxi;  
crse=sqrt(vecdiag(rv));  
title 'cluster robust standard errors';  
print crse;
```
