Improved regression estimation of a multivariate relationship with population data on the bivariate relationship[1]

Mark S. Handcock
University of Washington, Seattle

Michael S. Rendall
RAND and Pennsylvania State University

Jacob E. Cheadle
Pennsylvania State University

Working Paper no. 36
Center for Statistics and the Social Sciences
University of Washington

September 23, 2003

[1] Mark S. Handcock is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322 (Email: handcock@stat.washington.edu; Web: www.stat.washington.edu/handcock). Michael S. Rendall is a demographer at the RAND Corporation, Santa Monica, CA and Adjunct Associate Professor of Sociology and Demography, Population Research Institute, the Penn State University. Jacob E. Cheadle is a doctoral student in Sociology and Demography at the Pennsylvania State University. This work was funded by grants by the National Institute of Child Health and Human Development to the first two authors (R01 HD043472 01), to the Penn State University Population Research Institute (Core Grant R24 HD41025 and for Interdisciplinary Training in Demography 5T32 HD07514), and to the Center for Studies in Demography and Ecology at the University of Washington (Core Grant R24 HD42828). We are very grateful for programming assistance from Leslie Benson, and for comments from participants at seminars at Penn State University and the University of Washington.

Abstract

Regression coefficients specify the partial effect of a regressor on the dependent variable. Sometimes the bivariate, or limited multivariate relationship of that regressor variable with the dependent variable is known from population-level data. We show here such population-level data can be used to reduce variance and bias about estimates of those regression coefficients from sample survey data. The method of constrained MLE is used to achieve these improvements. Its statistical properties are first described. The method constrains the weighted sum of all the covariate-specific associations (partial effects) of the regressors on the dependent variable to equal the overall population association of one or more regressors. We refer to those regressors whose bivariate or limited multivariate relationships with the dependent variable are constrained by population data as being "directly constrained." Our study investigates the improvements in the estimation of directly-constrained variables, and also the improvements in the estimation of other regressor variables that may be correlated with the directly-constrained variables, and thus "indirectly-constrained" by the population data. The example application is to the marital fertility of black versus white women. The difference between white and black women's rates of marital fertility, available from population-level data, gives the overall association of race with fertility. The constrained MLE that uses this information both provides a far more powerful statistical test of the partial effect of being black, and purges the test of a bias that would otherwise distort the estimated magnitude of this effect. We find only trivial reductions, however, in the standard errors of the parameters for indirectly-constrained regressors.

Keywords: minority group hypothesis; combining data-sets.

1. Introduction

A typical situation in the social sciences is that data are available in aggregate levels of a societal unit, but have too little detail. These data are therefore considered useful only at a level of preliminary description or for cross-societal comparison. The best-known and most general of societal data collections in the United States is the decennial census of households and individuals. A glance at national or subnational statistical yearbooks, however, reveals many other population collections or registers of events of potential interest to sociologists. Individuals, businesses, and non-profit institutions register their coming into being, their ceasing to exist, and certain changes to their status. Annual income is recorded in corporate and individual tax returns. Entry to, and exit from, governmental support programs (low-income benefit programs, unemployment insurance, old-age and disability insurance, etc.) and punishment or supervision programs (prisons, probation, restraining orders, child-support orders) at the national, state, and local levels are typically registered and compiled. Limited amounts of information about the individuals experiencing the events are typically also collected.

Standard regression modeling techniques are not appropriately used with these population-level data, as there are too few variables to estimate behavioral models. Putting this into a statistical framework, sociologists employing regression methods are typically interested in studying the association between an dependent variable and an explanatory variable of interest (a "target variable"). The estimates of this association are of more interest after controlling for confounding associations of variables related to both the dependent variable and the explanatory variable of interest ("control variables"). Not infrequently, data at the population level are available for the dependent variable and the target variable, but not for an adequate set of control variables. In this case, only a "misspecified" model may be estimated with these population data. The sociologist then turns to survey data to specify the model more fully. Typically she or he then abandons the information on the overall associations between the dependent and target variable provided by the population data.

That social scientists are forced to choose *either* population data *or* sample data, however, need not be assumed. The advantages and disadvantages of sample versus population data collections point strongly to their being complementary. The advantages of population data are that they are without sampling error and may be much less subject to biases due to non-response. These advantages mirror the main disadvantages of survey samples: sampling error and bias due to non-response. The main advantage of sample surveys is that a large amount of information is collected about individuals, often with special thought in the selection of variables towards those that may have causal associations. These are the variables from which a behavioral model may be specified. The challenge then is to develop and apply statistical methods that combine population data's more precise and less biased estimates of the overall associations between the dependent and target variables with survey data's breakdown of these overall associations into multivariate associations between dependent, target, and control variables. The main purpose of the present study is to describe and illustrate a statistical method that combines the respective advantages of survey and population data. The method is constrained maximum likelihood estimation. Population-level data provide the constraints, in the form of information about the overall association between an dependent variable and a target explanatory variable. Estimates of the multivariate associations between the dependent variable and the predictor variables using survey data are then constrained to equal the overall associations.

Statistical Theory of Constrained Maximum Likelihood Estimation and Other Methods for Combining Population and Survey Data in Regression Analysis

The traditional statistical method for combining survey and aggregated population data is post-stratification (Kish 1965, Lohr 1999). Broadly defined, post-stratification refers to methods for adjusting bias and reducing variance in survey results by reweighting observations after selection (Smith 1991). Until recently, the statistical literature on the statistical properties of post-stratified estimates has been relatively sparse (Holt and Smith 1979), but has been applied also to the modeling of survey data using constraining information from population data (Deming and Stephan 1942, Ireland and Kullback 1968). More recently, Qin and Lawless (1994) develop methods for combining information from multiple sources when the information about the parameters of interest can be expressed in terms of unbiased estimating equations. They establish some theoretical properties of contingency tables with population data for marginal probabilities. Their equations are applied in conjunction with the empirical likelihood for the sampled information to estimate the parameters. This approach is closely allied to the econometric approaches we discuss below.

In part the lack of development of the post-stratification approach to modeling was because it has been largely motivated by, and is usually applied in, design-based inference. However, the ubiquitous nature of non-response and attrition has lead many researchers to instead consider model-based inference. In design (i.e. randomization) based inference, the population values are regarded as fixed numbers and inference is based on the probability sampling scheme for the survey. If non-response exists, design-based inference is inappropriate except under very strict assumptions about the nature of the non-response. In the alternative, model-based perspective, non-response and attrition are regarded as further sample selection, and the ultimate sampling scheme must be inferred from a selection model and the observed data. In this case post-stratification can be used to make the non-response ignorable for inference in the sense of Little and Rubin (1987). From this perspective the combination of survey and population data should be model-based (Little 1991, 1993, 1995). Little and Wu (1991) compare the design-based and model-based approaches when the sampled population differs from the population data due to non-response or coverage errors.

Little (1993) develops a Bayesian model-based approach to combine information from sample surveys and population data. The approach assumes the population distribution of a categorical variable in a simple random sampled sample survey is determined from the population information. His approach is to post-stratify on the variable using a Gaussian model for the response given the post-stratification variables. This approach is useful also when the survey is subject to simple forms of non-response and coverage errors. Further, the model is easily extended to the case where a joint distribution of multiple variables is available from the population data. Application to multiple datasets, some with fewer covariates, is also possible. This is analogous to population level data with fewer variables combined with survey data with more variables. This may be restated as a partially-missing-regressors problem (Little 1992). Including observations with some missing regressors nevertheless increases the efficiency of the estimation of parameters for the regressors that are present. In this way, missing data observations are treated as a second type of sample that must be combined with the sample of complete data observations. This framework can be extended to cover situations in which samples are combined from different studies with overlapping, but not completely identical, sets of regressors. Gelman and Little (1998) and Gelman and Carlin (2001) develop further method

of adjustment that are based on modeling population structure. Bethlehem (2002) shows how the Dutch POLS social survey can be adjusted for unit non-response, by including population level data.

There is a well developed literature on the econometric theory for combining population-level and sample-survey data. This work dates back to at least Manski and Lerman (1977). The approach is conceptually closely related to what econometricians refer to as choice-based, or endogenously stratified random sampling. In (bio)statistics this is often studied under the title of case-control or retrospective sampling schemes (Prentice and Pyke 1979, Breslow and Day 1980). Of most direct importance is the work by Imbens and Lancaster (1994, 1996) and Hellerstein and Imbens (1999). This is a very active research area in econometrics that has direct relevance to other social science fields. Imbens and colleagues (Imbens and Lancaster 1994; Hellerstein and Imbens 1999) explore the benefits of combining population with survey data, using economic data in a Generalized Method of Moments (GMM) regression framework. Imbens and Lancaster (1994) consider the estimation of parameters in the regression model under moment restrictions on the data contributed by population data, and report large gains in efficiency by incorporating marginal moments from census data with sample-survey joint distributions. Hellerstein and Imbens (1999) give an example of the bias-reduction possibilities of including aggregate data in the context of estimating wages from survey data. In that case, the survey data suffers from possibly non-random attrition. They also show how this may be addressed by means of a re-weighting scheme under an implicit assumption that the values are missing at random. This approach can be seen as an extension of post-stratification using a special case of the empirical likelihood estimator (Qin and Lawless 1994, Imbens, Johnson and Spady 1998). While the approach does not require parametric assumptions about the error distributions, it does not benefit from this information either. This reduces small-sample statistical efficiency and excludes Bayesian extensions relaxing the assumption that the constraint values are exact. In addition, the approach does not adjust for non-response which is not missing at random. Likelihood-based methods and maximum likelihood estimators in particular are more frequently found in sociological and much other social-scientific modeling work, due to their good statistical properties and because likelihood-based methods provide a general conceptual and inferential framework.

There is also a tradition of combining population and sample data in macro-level demographic analyses, using techniques including ‘model’ life tables and indirect standardization (Smith 1992). More recently, Handcock, Huovilainen, and Rendall (2000) demonstrated the potential of a constrained maximum likelihood estimator to combine sample survey data with birth registration data in the estimation of a multivariate model of fertility. Large gains in efficiency were achieved through the intercept term of their single-covariate logistic regression equation. The variance about the intercept parameter was halved when the General Fertility Rate constraint was introduced. Since the covariate-specific probabilities are functions of the intercept parameters, the reduction in variance in the constrained model was similarly large (around 50%) for the covariate-specific birth probabilities. Handcock et al did not, however, use any population-level information on fertility rates by model covariates, and so gains were confined to the intercept parameter and functions of it. The present article extends the Handcock et al results to consider possible gains in efficiency and unbiasedness additionally for the *coefficient* parameters of regression equations. In Section 2 immediately below, we describe the statistical theory of variance and bias reduction for the constrained maximum likelihood estimator, with particular application to logistic regression. In Section 3, we describe how this estimator may be used in a sociological application that combines panel survey data with

population data on black and white marital fertility in a test of the minority-group status hypothesis of fertility. In Section 4 the results are presented, followed by discussion in Section 5.

2. Statistical Theory

In this section we outline the statistical principles of variance and bias reduction when combining survey and population data in a likelihood framework. The exposition is oriented towards the type of estimators and assumptions about the nature of the survey and population data of our subsequent empirical application. We describe the theory of constrained MLE for survey data that are subject to varying degrees and types of non-response and attrition when exact population-level data are available to constrain the regression estimates. Further, we describe implementation with constraints on the weighted sum of conditional probabilities when the weights may be known from either population or survey data or both, and when the unconditional probabilities are known from population-level data. The basic principles also apply to non-likelihood-based regression methods such as least squares and method of moment estimators. We make reference to these generalities in the course of the exposition.

2.1 Representative Survey Data Combined with Exact Population Values

In this section, we consider what improvements could be achieved by combining survey data that are representative of our target population with accurate population-level data. To say that the survey data are 'representative' corresponds in the terminology of missing data to "ignorable" non-response (Rubin 1976), encompassing as special cases standard survey designs where data are 'non-missing', 'missing completely at random', 'missing at random', or subject to 'covariate only missingness'. We say that the population data are 'accurate' to mean that they are without substantial non-sampling error such as undercounting or misclassification. Because they are collected for all members of a target demographic population, we assume that they are statistically precise (that is, not subject to sampling error). In this section we show how the population values may be used to reduce sampling variance about survey estimates. We give a formula (equation 8 below) for the reduction in variances about the regression parameters, and that demonstrates that the standard errors of the estimates when using the population information will always be lower than when this information is ignored. This formula applies to the reduction in standard errors on both the intercept and coefficient parameters in the regression.

We start by writing the joint distribution of a response Y and covariates X as

$$P(Y = y,\ X = x \mid \theta_0) = P(Y = y \mid X = x,\ \theta_0)\, P(X = x), \qquad (1)$$

where θ_0 is the unknown parameter vector of interest describing the relationship between Y and X. The covariates X include the target and control variables. These may, for example, be the regression parameters in a logistic or probit regression of Y on X. Suppose the universe consists of women within a given range of childbearing ages, and that the response variable Y has two levels: 0 denotes no birth, and 1 denotes a birth, during the year $[t\text{-}1, t)$. Suppose further that the only covariate in this model is a dichotomous "pre-marital children" variable X for whether there are any children from before this marriage in the family unit at time $t-1$. The binomial logistic regression model for the birth probability $P(Y=1 \mid X = x, \theta_0)$ is given by:

$$\text{logit}[P(Y=1 \mid X = x, \theta_0)] = \theta_1 + \theta_2 x. \qquad (2)$$

Here the parameter is $\theta_0 = (\theta_1, \theta_2)$. Denote the survey data by $D = (y_i, x_i)$, i=1, ..., n. If this is all the information we have, under standard regularity conditions, the value of θ that maximizes the likelihood

$$L(\theta; y, x) = \prod_{i=1}^{n} P(Y = y_i,\ X = x_i \mid \theta) = \prod_{i=1}^{n} P(Y = y_i \mid X = x_i,\ \theta)\ P(X = x_i) \tag{3}$$

is an asymptotically efficient estimator of θ_0. Under these conditions, the estimator is also asymptotically unbiased and Gaussian with asymptotic variance V_s, where V_s is the inverse of $E_{\theta_0}\left[\partial \log[L(\theta; y \mid x)] / \partial \theta_{ij}\right]$, the Fisher information matrix for θ (Rice 1995). Note that as the sampled distribution of the covariates does not depend on θ_0 the same estimator is produced by maximizing the constrained likelihood

$$L(\theta; y \mid x) = \prod_{i=1}^{n} P(Y = y_i \mid X = x_i,\ \theta). \tag{4}$$

Intuitively this means that information about the population distribution of the covariates does not effect the estimator. Thus population information about a function of X alone would not improve the estimation of θ_0.

Now suppose that we supplement the survey data by population information about some function of the response and covariate variables, which we denote by $g(y, x)$. The function may be univariate, bivariate, or multivariate if information about multiple characteristics is available. As the information tells us something about Y or how Y and X relate to each other we might expect that it will also help us infer the value of θ_0. We assume that the information can be expressed as a mean of a multi-dimensional function over the population:

$$C(\theta_0) = E_{\theta_0}[g(Y, X)] \tag{5}$$

where the value of $C(\theta_0)$ is known from the population data to be ϕ, say. Most information can be expressed in this form by a judicious choice of $g(y, x)$. Returning to the example, consider the above model for birth probabilities in terms of presence of pre-marital children. Population data supply the annual probability of childbearing among all married couples in that wife's age group, ϕ. Survey data provide the proportion of couples with and without pre-marital children, $\rho = P(X = 1)$ and $1 - \rho = P(X = 0)$. By choosing

$$g(y, x) = \begin{cases} 1 & y = 1 \\ 0 & y \neq 1 \end{cases},$$

and using the above expressions for the covariate-specific birth probabilities, the constraint (5) is:

$$P(Y = 1 \mid X = 0, \theta_0)(1 - \rho) + P(Y = 1 \mid X = 1, \theta_0)\rho = \phi \tag{6}$$

In general, a constraint for covariates *x* and dependent variable *y* is of the form:

$$C(\theta) = E_{\theta}[g(Y, X)] = \iint_{x\,y} g(y, x)\ P(Y = y \mid X = x,\ \theta) P(X = x) dy dx \tag{7}$$

If the marginal distribution of the covariate X, $P(X = x)$, is known then this constraint is a known function of θ. Hence (5) constrains this function of θ to equal its known population value ϕ. If we maximize the likelihood (4) subject to this constraint using the procedure described in Handcock, Huovilainen, and Rendall (2000), the estimator is still asymptotically efficient, unbiased and Gaussian. However, while the asymptotic variance in the unconstrained version is given by the Fisher information matrix V_S, in the constrained version, the asymptotic variance is:

$$V_C = V_S - V_S H^T [HV_S H^T]^{-1} HV_S \qquad (8)$$

where $H = \left[\partial C_i(\theta) / \partial \theta_j\right]$ is the gradient matrix of $C(\theta)$ with respect to θ. As the second term in this expression is positive definite, the inclusion of the population information always leads to an improvement in the estimation of θ_0. In particular, the standard error of the estimator in the version using the population information (the constrained model) will always be less than the one that ignores it (the unconstrained model). A further result of (8) is that the asymptotic ratio of the variances of the constrained to unconstrained parameters is independent of the survey sample size. Thus, the *percentage* reduction in the standard errors of the regression parameters will be approximately the same for all sample sizes.

Importantly also, both V_S and H in (8) can be estimated from the survey data using the unconstrained model. Then the efficiency gain from including population information can be estimated before running a constrained model, and so before obtaining the population data. Alternative choices for $g(y,x)$ can then be compared in terms of their statistical efficiency and ease of collection of the population information. Note that the increase in efficiency from including population data will be reduced if the population distribution of the covariates is not known. Typically in demographic applications, the population data will provide information about the univariate or bivariate distributions of at least some of the covariates, but survey data will be needed to provide information about the multivariate dimensions of the covariate vector.

[Figure 1 about here]

The effect of including the population data can be understood graphically in Figure 1. This is a stylized representation of the relative positions in a two-dimensional parameter space of the various models. The target population model is represented by θ_0. The dashed curves are the contours of the unconstrained likelihood (3) with the maximum likelihood estimator using only the sample survey represented by θ_{MLE}. The models satisfying the constraints (5) on the parameters imposed by the population data are represented by the thick line. Note that θ_0 is always on this line if the population data are from the target population. The constrained MLE θ_{CMLE} is the value on this line that maximizes the likelihood subject to the constraints. The variance formulas given above show that θ_{CMLE} is, on average, closer to θ_0 than θ_{MLE}. The exact size of the improvement depends on constraints, but is calculable from (8) for given $g(y,x)$, independently of the actual value of the constraint ϕ. Because the survey data are representative, though, the expected values of both the constrained and unconstrained parameters are equal to the population parameter θ_0. The gains realized through the introduction of the population data

will be in variance reduction only, as the unconstrained estimator is already unbiased with respect to the target population.

It is worthwhile to be explicit about the situation where the survey suffers from non-response but is still assumed to be representative. The simplest case is where the values of D that are missing are a simple random sample from the complete sample This is known as the *missing completely at random* condition (Little and Rubin 1987). Denote the observed part of D by D_{obs}, and the missing part by D_{mis} so that D=(D_{obs}, D_{mis}). As the mechanism that determines which values are missing is independent of the response, covariates and θ_o, the likelihood for the data is just the product (3) taken over the observed cases. In addition, the constrained MLE is still asymptotically unbiased, efficient and Gaussian. The asymptotically efficiency for the observed data relative to the complete data is just the proportion of the complete data that is missing. Hence the inclusion of the population data has the same *rate* of increase in efficiency is it has in the complete data case.

This assumption can be weakened to allow values to be missing differentially by covariates, but independent of the response and of the vector of parameters θ_0 relating the response variable to the regressors. We call this *covariate only non-response* (Rubin 1977). In our example, this would mean that childless women might have a different response rate to women who had at least one child at time $t-1$, as long as the rate was the same of all women in each category regardless of their birth status during the year. That is, the reason for non-response is then perfectly related to the measured covariates. As the constrained likelihood is expressed in terms of the conditional distributions $P(Y = y_i \mid X = x_i, \theta)$, and these are unchanged, the likelihood for the data is again just the product (3) taken over the observed cases. Thus even though the observed sample distributions of the covariates are biased, the inclusion of the population information is unaffected and the constrained MLE is asymptotically unbiased, efficient and Gaussian. The asymptotic efficiency for the observed data relative to the complete data is again just the proportion of the complete data cases that is missing.

Suppose now that the non-response also depends on the response Y. Suppose also that the probability that an observation is missing may depend on D_{obs} but not on the missing part D_{mis}. This is known as *missing at random*, in the sense of Rubin (1976). This is less restrictive than covariate only non-response, which is in turn less restrictive than missing completely at random. Missing at random allows the probability that a datum is missing to depend on the response and covariate of the datum itself, but only indirectly through the quantities that are observed. Let us assume that the parameters of the non-response mechanism are distinct from the parameters of interest θ_o. If both the missing at random and the distinctness conditions hold, then the missing-data mechanism is said to be *ignorable* (Little and Rubin 1987). This is an important concept, because we can still model an ignorable non-response mechanism as a function of what we have observed and obtain efficient inference for θ_0. Finally note that this assumption, like all those made about missing data, can not be verified from the observed data alone; support must come from expertise or information external to the data. For example, we might know from other studies that women's marital fertility by pre-marital fertility does not differ by whether they respond to surveys, even if, say, women with no pre-marital fertility are less likely to respond. If there is non-response, the observed information includes not only the observed values of the response and covariates, but also a variable R indicating if the case was observed or not. Hence the likelihood for the observed data is the joint likelihood for D_{obs} and R. However, if the missing data mechanism is ignorable, the joint likelihood is:

$$L(\theta; y_{obs}, x_{obs}, R) = P(R \mid Y = y_{obs}, X = x_{obs})\ P(Y = y_{obs} \mid X = x_{obs}, \theta)$$
$$= P(R \mid Y = y_{obs}, X = x_{obs})\ L(\theta; y_{obs} \mid x_{obs}). \quad (9)$$

The first term is independent of θ so the likelihood is proportional to the constrained likelihood for the observed values of the response and covariates. Hence the MLE is just the constrained MLE under $L(\theta; y_{obs} \mid x_{obs})$, independent of the missing data mechanism. That is, under our approach we can achieve fully efficient inference without modeling the missing data mechanism explicitly for likelihood-based inference about θ_o. The constrained MLE under $L(\theta; y_{obs} \mid x_{obs})$ is asymptotically unbiased, efficient and Gaussian. The asymptotical efficiency for the observed data relative to the complete data is just the proportion of the complete data that is missing. Hence again the inclusion of the population data has the same *rate* of increase in efficiency as it has in the complete data case.

2.2 Non-representative Survey Data Combined with Exact Population Values

In this section, we consider what improvements could be achieved using accurate population-level data to supplement 'non-representative' survey data, meaning survey data that are less than perfectly representative of the population on at least one dimension related to the estimation problem. In missing data terminology, the effective survey sampling mechanism then incorporates non-response that is no longer ignorable. The statistical properties of estimators that combine non-representative survey data with exact population values, however, apply more broadly than to the case of 'non-ignorable' survey non-response. They can be extended also to respondent misreporting (Schafer 1997) and to survey sampling designs that do not match exactly to the target population.

The main result of using constraints from the target population in combination with the above kinds of survey data is that the more population information we introduce, in the form of constraints about the relationship between *Y* and *X*, the closer we will get to unbiased regression estimates of target population relationships. We describe this below in terms of the synthetic population that is formed by the combination of elements from both the representative and non-representative effective sampling frames respectively from the population-level and survey data. The inclusion of population constraints moves this synthetic population towards the target population. The more constraints used, the closer is the synthetic population to the target population, and the greater the reductions in bias about parameters estimated from this synthetic population.

[Figure 2 about here]

Formally, following Little and Wu (1991) and Hellerstein and Imbens (1999), we refer to the survey data's distribution as the *sampled population*, with parameter θ_{sample}. We distinguish this from the *target population*, with parameter θ_o. The maximum likelihood estimator will approach θ_{sample}. If the survey is representative, then $\theta_{sample} = \theta_o$; otherwise the difference between them represents the bias of the sample survey. The inclusion of population information will, in general, reduce this bias. Suppose we use the constrained MLE to estimate θ_o. Consider the synthetic population which satisfies the constraints (5) defined by the population data and that is closest to θ_{sample} in terms of likelihood (see Figure 2). This population is in a sense a combination

of the sampled population and the target population. We denote the parameter for this synthetic population by $\theta_{combined}$. The constrained MLE θ_{CMLE} will approach $\theta_{combined}$, as the sample size increases, rather than the true value θ_o. Thus the difference between $\theta_{combined}$ and θ_o is a measure of the bias that remains after introducing population constraints. In this sense it is the *bias of the combined survey and population information*. In general $\theta_{combined}$ will be closer to the true value θ_o than θ_{sample}, so the inclusion of the population data improves the estimates. The development of the properties of the constrained MLE in Section 2.1 still applies, with θ_o now replaced by $\theta_{combined}$. In particular, the variance formulas given above now apply for θ_{CMLE}, which is, on average, closer to θ_o than is θ_{MLE}. Hellerstein and Imbens (1999) derive these results in the special case of linear regression.

Computational limitations for the maximization problem may be encountered when many population constraints are simultaneously applied. To circumvent this problem, the post-stratification reweighting approach of Hellerstein and Imbens can be used. Those authors demonstrate the equivalence between the constrained estimation and the reweighting approaches in the linear regression context. They show that the empirical likelihood MLE can be expressed as a weighted linear regression estimator, where the weights are a by-product of calculating the MLE, and can be interpreted as post-stratification weights. This approach has the major advantage that once the weights are calculated, the weighted dataset can be used within standard statistical packages, and interpreted accordingly. Both the constrained methods and the reweighting methods lead to identical estimators. The choice of method can then be made on the basis of ease of programming and statistical computational efficiency.

Misreporting of respondents can be statistically treated in a similar way to non-ignorable non-response. It requires the modeling of the misreporting mechanism just as the non-response is modeled. Even parameters of the misreporting (i.e., the misreporting probability) can be included as parameters. A good example is Heitjan and Rubin's (1989) treatment of respondents rounding of answers (e.g. age, income). See also Heitjan (1990) for a review of these methods.

2.3 Application to logistic regression

Let Y be a binary response variable modeled via a logistic regression on covariates $X = \{X_1, X_{2,}..., X_q\}$:

$$\text{logit}[P(Y=1 \mid X_1 = x_1, X_2 = x_2..., X_q = x_q, \theta_0)] = \sum_{k=1}^{q} \theta_{0k} x_k. \qquad (10)$$

The responses are assumed to be conditionally independent given the covariates. We next introduce constraints on the total effects of each variable. Let ϕ_{ij} be the proportion of positive responses *(Y=1)* in the population with $X_i = j$. The corresponding constraint functions $C_{ij}(\theta)$ are each of the form:

$$\phi_{ij} = C_{ij}(\theta) = E_\theta[g_{ij}(Y,X)] = \sum_{x:x_i=j} P(Y=1 \mid X=x, \theta)\, \pi(X=x \mid X_i = j) \qquad (11)$$

where

$$g_{ij}(y,x) = \begin{cases} 1 & y=1 \text{ and } x_i = j \\ 0 & \text{otherwise} \end{cases}$$

The constraint function (11) is expressed as the sum over of two product terms (on the right hand side of the constraint function). The first term is the probability of a positive response conditional on the value of the regressor vector *x*. The second term is the proportion of the population with a specific set of values on the regressor variables given that $X_i = j$.

Consider now the case in which there is a constraint function defined for each possible value *j* that variable X_i may take. For example, X_1 may be a "race" variable that takes the value of $X_1 = 0$ for whites and $X_1 = 1$ for blacks. Continuing this example, let response variable *Y* indicate whether a birth occurs in the year. Then $\{\phi_{11}, \phi_{10}\}$ represents the bivariate association of race with fertility in the population of whites and blacks. For example, ϕ_{11} / ϕ_{10} expresses the ratio of black to white fertility. The two constraint functions $C(\theta_{11})$ and $C(\theta_{10})$ then together constrain the marginal effects in the behavioral model to preserve this overall ratio of black to white fertility that is known from the population data. When a set of constraints of this type is applied, so that for a given regression variable X_i its bivariate association with *Y* is completely specified for the population, we say that the regression variable X_i has been *directly constrained*. For other variables in the equation, we say they are *indirectly constrained*, since each constraint potentially influences every parameter value via the interrelationships between the covariates. Hence a direct constraint on one variable indirectly influences the values of the other parameters. These indirect effects are expected to be larger if the covariates are closely related. For example, if a variable is an interaction between a directly constrained and another variable, it may be significantly influenced even if it is not directly constrained itself. The magnitude of any resulting reductions in variance and bias is one of the subjects of our empirical investigation reported in the results section below.

In general, the population proportions $\pi(X = x \mid \mathrm{X_i} = j)$ may be estimated from survey or population data, or a combination of the two. Estimating them from the survey data is useful to ensure conformity of the variable definitions between the first and second terms on the right hand side of the constraint equation, but does introduce a source of variability in the constraint function. It is possible to modify equation (8) to include a component due to this uncertainty. The constrained estimator is still used, but with the constraint functions conditional on the distribution of the regressors estimated from survey data. Let $g(x;\theta) = E_{Y|X=x;\theta}[g(Y,x)]$ be the conditional expectation of $g(Y,X)$ given $X = x$. The variation of $g(X;\theta)$ determines the variation in the constraints. The variance of the constrained estimator is given by Imbens and Lancaster (1994):

$$\mathrm{V_{CU}} = [\mathrm{V_S^{-1}} + \mathrm{H^T}\Delta_g^{-1}\mathrm{H}]^{-1} = \mathrm{V_S} - \mathrm{V_S H^T}[\mathrm{HV_S H^T} + \Delta_g]^{-1}\mathrm{HV}$$

where Δ_g is the covariance matrix of $g(X;\theta_0)$. Compared to equation (8), it can be seen that Δ_g represents the cost of not knowing $\pi(X = x \mid \mathrm{X_i} = j)$ and this inflates the variance of the estimator. When Δ_g is close to zero, that is the constraint right hand side varies little from sample to sample, the two formulas are very close. The effect of using sample data to estimate population proportions $\pi(X = x \mid \mathrm{X_i} = j)$, however, is to unambiguously reduce the effectiveness of constrained estimation. By just how much will vary from application to application, but can be quantified for known or hypothetical Δ_g via this formula.

3. A Constrained MLE Test of the Minority-Group Status Hypothesis of Fertility

The example introduced above in the context of the theoretical properties of constrained maximum likelihood estimators can now be elaborated and estimated. Specifically, we consider a test of the "minority-group status" hypothesis of couple fertility. We demonstrate the constrained maximum likelihood method by testing this hypothesis on the marital fertility of black and white couples in which the wife is aged between 30 and 34 years old. According to the minority-group status hypothesis, fertility will be lower for a minority group of otherwise equal economic status (Goldscheider and Uhlenberg 1969). In previous tests of the hypothesis applied to minority black versus majority white women (Johnson 1988; Boyd 1994), the strength of empirical support for the hypothesis has varied. Our test of the minority group hypothesis is confined to the 30 to 34 year-old age group for simplicity of statistical estimation and testing. While limiting the age range in this way will mix childbearing quantity with timing effects, we include regressors that control for timing associations.

The minority-group status theory of fertility differentials was originally proposed when fertility within marriage predominated among both majority and minority groups. Thus it is appropriate for us to apply the hypothesis to marital fertility. Johnson's finding that higher educated black women's lower fertility may be partly due to their higher rates of divorce also argues for separate consideration of marital fertility. It is important, however, to take into account the potentially confounding effect of fertility before the marriage began. This may be non-marital fertility, either with her current husband or with a previous partner. It may also be fertility within a previous marriage. In either case, we expect and find that this depresses fertility in the current marriage. Because, as we show, pre-marital fertility is substantially more common among black married women than among white married women, it is important to include a control for any pre-marital children in addition to other socio-demographic and economic variables.

We obtain the bivariate associations between race and fertility from the population data: the race- and year-specific marital fertility rates for 30 to 34 year old women from 1984 to 1993. The population data are the National Center for Health Statistics' (NCHS 1999) published estimates of annual marital fertility rates by five-year age group and race of the woman. The NCHS makes these estimates by using as their numerator, all marital births in the given year 1984 to 1993 to married mothers of that racial group[2], and by using as their denominator the Census Bureau's mid-year population estimates by sex, age, and marital status. We assume the marital age-, period- and race-specific fertility rates have zero sampling variance, and that the age, race, and period definitions are those of the target population for our analyses. That is, we assume the NCHS data represent the true population values exactly.

In general, when jointly using survey and population data, there will seldom be an exact match between the population and survey universes and variable definitions. In the present study, we specify the population about which we wish to make inferences on the basis of the population data, and use the survey data to best approximate that. This allows us to explore empirically the statistical properties of the case described above in terms of statistical theory. This is the case in which the population values are known exactly and the sample may or may not have been drawn (or have subsequently evolved) in a way that can be said to exactly represent that target population.

[2] In general, these are 100% samples, but some states provide 50% samples for the national compilation by NCHS, who then weight them to the population total.

The survey data we use are from the Panel Study of Income Dynamics (PSID, Hill 1992). These data have the advantages of coming from a very long-running panel survey with covariates for socio-economic and demographic variables available for each year. For the present study, it is especially advantageous to have economic status measured directly each year with an income variable. We make estimates for the years 1984 to 1993. The latter year is the most recent year for which "final release" files, including generated variables such as the family income-to-poverty ratio, were available from the Survey Research Center when we coded our survey data. The PSID sample uses an unequal probability sample design. We account for this by conducting our estimation with the PSID's individual sample weights. Thus we account for both initial sample design effects and some of the biases introduced by attrition. The PSID's sampled population may, however, have drifted away from target population universe over time, through differential attrition (Fitzgerald, Gottschalk, and Moffitt 1998), and through its not capturing the processes of population change through immigration. We interpret such drift in terms of *bias* with respect to the target population.

The universe consists of years of exposure to marital fertility in the period 1984 to 1993 among white and black married couples in which the wife is aged 30 to 34 years old. The survey data were collected at annual intervals late in the year, with age recorded in completed years at last birthday. Using these data to best approximate our calendar-year universe, we select our sample to consist of all survey year-pairs $(t\text{-}1,\ t]$ in which the wife was aged from 29 to 34 in year $t-1$, and hence aged 30 to 35 in year t. The part-year exposure while still aged 29 is then balanced by part-year exposure at age 35. The matching of survey period to calendar year is done at year t of the $(t\text{-}1,\ t]$ survey period, since the survey data are collected late in each year, and thus the majority of exposure in each $(t\text{-}1,\ t]$ year occurs during calendar year t. Married couples each contribute up to six couple-years of exposure in the 1984 to 1993 period. This results in a sample of 8,266 person-years. We ignore variance-estimation complications due to the repeated observation of individuals in the panel. Because we use the same data for both the constrained and unconstrained estimates, introducing this further complication should not change our main results.

The variables from the PSID that are used in the regression are as follows. The dependent variable Y has two levels: 0 denotes no birth, and 1 denotes a birth to the couple, during the year $(t\text{-}1,\ t]$. Using the PSID's panel data only, we code a birth when a child aged less than two has entered the family unit since the previous year. This assumes that the parents and child live together at the survey interview immediately following the birth, and that infant mortality between birth and survey is zero. Improvements to the accuracy of coding of births could presumably be achieved through supplementing the panel data with the PSID's fertility histories. For the present study's methodological objectives, though, the panel data are sufficient.

Next, we consider the explanatory variables. A fully-specified model of the determinants of the marital birth event would include a variety of demographic, economic, and sociological variables influencing the probability of a birth in the year. The demographic variables might include the single-year ages of both the wife and the husband, length of the marriage, number of, and years since, previous births within this marriage, for this couple (allowing for children born to the couple before they married), and outside this couple. The economic variables might include the husband's and the wife's current and opportunity wages (the latter affected by education and years of working experience), and the couple's net worth. Sociological variables, such as religious affiliation and behavior, and attitudes towards marriage and family, might also be included. Our intention in this example is to include a reduced set of these variables that is

sufficient to illustrate the statistical issues and advantages when using constrained MLE as opposed to the usual, unconstrained technique.

As proxies for a full set of demographic variables, we include single-year durations of marriage up to 10 years and over, and whether the age of any child in the current family unit is greater than the duration of the marriage (a "pre-marriage children" variable). As a proxy for a full set of economic variables, we include dummies for the family's tercile income-to-poverty ratio. This defines lower, middle, and upper-income married couples of this age group. The cut points are at 3.15 and 5.16 times the PSID's approximation of the official poverty ratio. Since the poverty ratio is determined by size of family, this also proxies partially for that demographic variable. As a proxy for sociological variables that have changed over time, including attitudes to marriage and the family, we include single-year period dummies for 1984 to 1993. These period dummies will, of course, proxy also for changes in other unobserved variables, including contemporary labor-market conditions.

Finally, race of the couple is coded from the "race of the Family Unit Head" variable in the PSID. A more complete model would allow for the race of the husband and wife to differ, but again, our simplification is adequate for illustrative purposes. Race is also interacted fully with each of the year dummies and with the dummy for whether there are any pre-marital children. The choice of these particular interactions is more for the purposes of illustration than for substantive or model-fitting reasons. The year dummies are interacted with race because the population data allows us to do so very precisely. The "pre-marital children" variable allows us to explore the effect of interacting a variable for which the population data do not provide direct information, with a variable (race) that is more directly constrained. We explore the effect of constraints on the black dummy and its interactions with pre-marital children and period, under different model specifications and sample sizes. A major focus of our study is the improvements of estimation and inference with respect to the directly-constrained coefficients for race and race-by-year interactions, as it is these that allow us to test the minority-group hypothesis of fertility.

Formally, the covariates in this model are represented by the vector x, measured at time $t-1$. The binomial logit model for the birth probability $P(Y=1|X=x, \theta_o)$ is given by equation (10). In the unconstrained case, we use the survey data alone to estimate the value of θ_0 that maximizes the unconstrained log-likelihood. We next introduce constraint functions. Let ϕ_{tr} be the NCHS fertility rate for black and white couples' ($r = 0, 1,$ respectively) in year $t = 1984, \ldots 1993.$ These fertility rates are used to constrain the black and white couples' annual probabilities of a marital birth. The 20 constraint functions $C_{tr}(\theta)$ are each of the form seen in equation (11) above:

$$\phi_{tr} \;=\; C_{tr}(\theta) \;=\; \sum_{x:year=t,\ race=r} P(Y = y \mid X = x, \theta)\ \pi(X = x \mid year = t,\ race = r) \qquad (12)$$

Thus the constraint functions are of the form in which the value of the dependent variable is known exactly for race and period subpopulations of the overall population of 30 to 34 year-old married women. The bivariate association of race with fertility is then implicitly constrained for all married 30-34 year-old women in any given year. The first term on the right hand side of the constraint function is the probability of a birth in year t conditional on the value of the regressor vector x. The second term is the proportion of the population of that race in year t with a specific set of values for the regressor variables. Here, it is the proportion of the population with a specific number of years marital duration, presence or absence of pre-marital children, and

family income-to-poverty tercile (that is, values on the regressors that are not directly constrained). We estimate those proportions from the survey data. In a simple specification of the model from our example problem, variance reduction in the estimation of our parameter interest (the coefficient on the black dummy) was little changed when variability about the regressor distribution was accounted for ----down from a 98.4% reduction over the unconstrained estimate's variance to a 97.0% reduction . Variance reduction in the intercept parameter, however, was more substantially affected----down from a 94.9% reduction to only a 73.2% reduction. For simplicity, the results presented below do not account for this source of variability in the estimates of variance reduction.

The maximum likelihood estimator under the constraints is the solution of:

$$\max_{\theta}[L(\theta; y \mid x)] \quad \text{subject to} \quad C_{tr}(\theta) = \phi_{tr} \quad t = 1,\ldots 10, \; r = 0,1$$

The estimator is asymptotically efficient, unbiased and Gaussian with covariance matrix approximated by (8). To evaluate the gains obtained by imposing constraints from the population data, we estimate both the unconstrained and constrained versions of our model, and compare the parameters and their standard errors.

We implement our constrained ML estimator using the PROC NLP procedure of the SAS/OR package (SAS Institute 1997). This procedure allows for a wide range of objective functions and for a large number of either linear or non-linear constraints. In the present, logistic regression case, the constraints are non-linear due to the non-linearity in the logistic cumulative density function. The NLP procedure also calculates the covariance matrix and standard errors, using the standard asymptotic approximation of the inverse of the Fisher information matrix. The NLP procedure is relatively simple to implement, and converges within a reasonable time (under two hours CPU time) for the specification presented here. In more complex specifications with larger survey samples, however, more flexible and efficient programming implementations may be required.[3]

4. Results

Comparisons between the population and survey estimates of the annual marital fertility for white and black married women aged 30 to 34 are shown in Figures 3 and 4 for the years 1984 to 1993[4]. Here, as throughout the analyses, the survey estimates are weighted. The degree of fluctuation due to sampling error is high for both races. This is seen both in the confidence intervals that are plotted with the point estimates, and in comparison with the population rates. The population rates show clear upward trends, and very little fluctuation from year to year, for both whites and blacks. The upward trends are not clearly visible in the highly-fluctuating survey rates. Several features stand out when comparing the white and black rates. First, the black fertility rates in the survey fluctuate more than the white fertility rates. This is as expected given the smaller sample sizes of married black women----approximately 250 person-years of exposure

[3] The main programming disadvantage of the NLP procedure is that it does not allow for the specification of the constraint function in matrix form. Thus it becomes unwieldy when the number of possible regressor-vector values to sum over becomes large.

[4] Strictly, the annual probability of a birth is estimated in the PSID, but this is equivalent to an annual fertility rate due to the negligible mortality at the childbearing ages.

per year, compared to married white women's approximately 600 person-years of exposure per year.

Second, while the black survey rates fluctuate both above and below the population rates, the white survey rates appear to be systematically higher than the population rates. In each of the ten years, the white survey point estimate is above the population rate, and in six the white population rate lies below the lower limit of the survey estimate confidence interval. In only two years does the black population rate lie outside the survey rate confidence interval. Thus the white survey rates exhibit strong evidence of upward bias, while the black survey rates show little apparent bias. This finding is important for the statistical testing of the minority-group hypothesis, as it could induce a spurious finding of lower marital fertility among black women, one that is due to bias in the sample of white women.

Third, in both the population and survey rates, the marital fertility of blacks is substantially lower than that of whites. This is seen clearly in the population rates, in which the black rates are consistently lower by about 20 percent. The black rates increase from an annual rate of .0675 in 1984 to .0796 in 1993, while white rates increase from .0830 to .1004 in the same period. These differences are the population estimates of the bivariate association of race and fertility. While this direction of differences is consistent with the minority-group hypothesis, before drawing a conclusion it is necessary to control for socio-economic covariates that may depress black marital fertility rates relative to those of whites of the same age group.

[Figures 3 and 4 about here]

Overall, these results show that married black women have substantially lower fertility than do white married women of the same age group. For the minority-group hypothesis to be supported, however, we should find that black women's fertility is substantially lower *after* controlling for socio-demographic and economic variables. To do so, we allow the survey data to contribute information on pre-marital children, on duration of the current marriage, and on the economic condition of the family. After including these demographic and socio-economic control variables, the coefficient for the race variable together with dummies for the interaction of race by year provide a test of the minority-group hypothesis.

The distributions of blacks and whites on the covariates, and the proportions giving birth by covariate and race, are shown in Table 1. Black women are likely to have been married fewer years (29.1% fewer than 5 years, compared to 23.0% of white couples). They are much more likely to have a child present from before the marriage (19.7%, compared to only 7.5% of white couples). They are also much more likely to be in the lower income-to-poverty tercile (50.3%, compared to 31.1% of white couples), and much less likely to be in the upper income tercile (16.7%, compared to 35.3% of white couples).

Two of these racial covariate distributions point to black couples having lower birth probabilities than white couples, and one points in the reverse direction. Having a pre-marital child and being in lower income categories are associated with substantially lower probabilities of giving birth: 0.035 with a pre-marital child compared to 0.116 without a pre-marital child; and 0.074 in the lower income category compared to 0.152 in the higher income category. Birth probabilities are higher, however, at shorter marital durations (the third and fourth years being the peaks), with the exception of the first year of the marriage. The lower marital fertility rates of 30 to 34 year-old married black women that are seen in the population data, therefore, may be due to socio-economic covariate differences by race, and not to a "minority-group status" effect.

Hence, as proposed by the minority-group hypothesis, it is necessary to control for these covariates to assess whether there is a depressing effect of minority status on fertility.

[Table 1 about here]

We estimate the model with socio-economic covariates under two regression specifications, alternately including and excluding the income variables (see Table 2). In both models, we control for marriage duration and presence of a pre-marital child when estimating the year-by-year effects of being black on having a birth. For each specification, we estimate the models alternately in their unconstrained and constrained versions. Comparing the specifications alternately with and without the income-to-poverty variables, a test of increase in the log-likelihood reveal that in both the unconstrained and constrained versions, inclusion of the income-to-poverty variables improves the fit ($p<.001$). The values of the log-likelihood for each equation are given in Table 2. As the standard errors for these parameters are trivially reduced by introducing constraints, it follows that the test for improving the model fit is almost identical when comparing the two specifications in the constrained models versus in the unconstrained models.

Our main focus is on the differences between the unconstrained and constrained estimates. As described in the theory section, the parameters can be divided into those that are directly constrained and those that are not. The former category consists of the year variables, the black variable, and the black-by-year interactions. For each year, there are black and white constraints. The difference between them measures the bivariate association of race with fertility in that particular year. For each year, there is also a survey variable whose regression coefficient measures the *marginal effect* of being black in that particular year. Thus these variables fit our definition from the theory section of having directly-constrained parameters. The category of indirectly-constrained parameters consists of those for the pre-marital child, marital-duration, and income-to-poverty variables, plus the variable for the interaction of black and pre-marital child.

We consider first *efficiency* gains from constraining the survey estimates. Consistent with the statistical theory presented above, the standard errors on all parameters are lower in the constrained equation than in the corresponding unconstrained equation. For the variables that are directly constrained, the reductions in standard errors are extremely large. For the year dummies, the reductions are by factors of more than 15, while for the black-by-year interactions, the reductions are by factors of up to 30. The reductions in standard errors about the black variable are by factors of about 9. Accordingly, the magnitudes of the coefficients for the year dummies and black-by-year interactions vary much less over the 1985 to 1994 period in the constrained model than in the unconstrained model.

[Table 2 about here]

The reductions in standard errors are of trivial magnitudes, however, for all indirectly-constrained variables, in all cases by less than 1 percent. Noteworthy here is that even for the interaction variable between black and non-marital child, the standard error is reduced by less than 1 percent. This negligible reduction is in spite of the strong correlation between being black (a directly-constrained variable) and having a pre-marital child (seen above). This example

points to the likelihood that any reductions in the standard errors of other than directly constrained variables in sociological research will be very modest.[5]

The intercept parameter is a special category. It is subject to a direct constraint, that of the fertility rate of the reference white group in the reference year 1984. However, the intercept parameter measures more than just an implicit marginal effect of being white in 1984, measuring also the implicit marginal effects of being in the reference categories of the variables that are not directly constrained. The reduction is by "only" one-third about the intercept variable. The statistical interpretation here is that the intercept is related both to the directly constrained variable (black versus white) and the unconstrained variables (marital duration and pre-marital child and, in the second specification, also income level). The more indirectly-constrained parameters there are in the model, the less will the intercept term be determined by the value of the constraint on the reference year 1984 for reference race "white". Comparing the specifications alternately with and without the income-to-poverty variables is instructive here. The reduction in the standard error about the intercept term is proportionately smaller when the income-to-poverty variables are additionally included (by 25.4 percent, versus by 31.8 percent in the model without the income-to-poverty variables). There are not, however, any substantial differences in the directly-constrained parameters with the additional indirectly-constrained parameters included. This is an important result, because the present study has used a simpler specification of the socio-economic and demographic variables than would typically be used. The finding of almost equally large reductions in the standard errors when further indirectly-constrained regressors are added indicates that directly constraining a regressor of interest can also be expected to yield very large efficiency gains in a fully-specified regression model.

We now proceed to evaluate the effects of these reductions in standard errors on our statistical evaluation of the minority-group hypothesis. We do so for the model that includes the income-to-poverty dummies. Because we have interaction dummies for nine years as well as a black dummy, we can consider the minority hypothesis in each year. A measure of the black minority-status effect for 1984 is the coefficient of the black dummy. For 1985-94 we can use the sum of the coefficients of the black dummy and the black-by-year interaction dummy. Together there are 10 statistical tests of the minority-group hypothesis that may be performed in each equation. Each minority-group hypothesis can be evaluated using a *t*-test whose test statistic is constructed by dividing the minority-status effect estimate by its standard error.[6] We then calculate confidence intervals around the estimate for each year. These are shown in Figure 5. The unconstrained model estimates of the year-by-year black minority-status effect have large confidence intervals around them, such that in no year is the estimate significantly different from zero. While point estimates are relatively evenly spread between being above and below zero (six above zero and four below zero), those below zero are further from zero, suggesting a possible effect that the statistical test may not be powerful enough to detect. This suggestion is supported by the results for the constrained model. Seven out of ten years has a point estimate below zero, and six of these are statistically significant. Only one of the years (1984) sees a point estimate that is above zero and statistically significant.

[5] We further confirmed this finding in results not reported here in which we experimented with simulated data that we created to have very high correlations between directly-constrained and other variables.

[6] For 1985-94 the test statistics are the sum of two estimated coefficients and so the standard errors can be calculated from a quadratic form in the appropriate variance-covariance matrix. This is V_S in the unconstrained case and V_C in the constrained case.

[Figure 5 about here]

We next assess whether and by how much we have reduced bias by constraining the estimates. As discussed in Section 2.2, the survey data is unbiased if $\theta_{sample} = \theta_0$. We can test the overall hypothesis of equality between these two parameter vectors using a Wald test statistic:

$$(\theta_{MLE} - \theta_{CMLE})^T [V_S - V_C]^{-1} (\theta_{MLE} - \theta_{CMLE})$$

which is asymptotically Chi-squared with degrees of freedom equal to the number of constraints. For the model including the poverty variables the test statistic is 69.86. The *p*-value based on the Chi-squared on 20 degrees of freedom is less than 0.001. Hence we can reject the null hypothesis of no bias against the alternative of bias in at least one parameter. This result is unsurprising given the comparisons between the sample and population fertility rates shown above for whites in particular (Figure 3). In particular, it seemed that one reason that black marital fertility in the PSID might be lower than white fertility is that the white fertility might have been upwardly biased.

We calculate the bias as the log-odds of having a birth calculated from the constrained model estimates minus the log-odds of having a birth calculated from the constrained model estimates. This calculation of the log-odds is just the right hand side of equation (10). We calculate this bias for the reference category on the other regressors, that is, no pre-marital child, marital duration 4 years, and bottom tercile income-to-poverty ratio. We do this separately for whites and blacks (see Figures 6 and 7).

[Figures 6 and 7 about here]

Returning to the results presented in Figure 3, we saw that there appeared to be a tendency towards a downward effect on fertility in the unconstrained model, but one for which the year-by-year statistical tests were not powerful enough to detect. One way of increasing the power of the test is to consider an average of each of the annual effects. This permits an overall test of the minority-group status hypothesis for the 1984 to 1993 period. We calculated the average 'black' effect in the constrained and unconstrained models as the mean of the annual differences in the log-odds, and calculated the standard error about this mean.[7] The mean effects are then -.375 (0.165 standard error, *p*=.02), or a 0.687 odds-ratio for the unconstrained model, and -.128 (.0445 standard error, *p*=.004), or a 0.880 odds-ratio for the unconstrained model. These are very good summary measures of the effect of constraining the survey estimates of the marginal race effect to known population values of the overall race and fertility association. The unconstrained model indicates that black marital fertility in this age group is as much as one-third lower than white marital fertility (in terms of odds of having a birth in the next year), after controlling for socio-economic variables and variables for marital and fertility history. This difference is statistically significant at the .05 level, but not at the .01 level. Thus although the magnitude of the effect is apparently very high, the statistical test is only just powerful enough to detect it. The constrained model indicates that the effect of black minority status is much smaller ---- a 12 percent lower odds than for whites of having a birth in the next year. The test for the minority status effect, however, is a very powerful one. The standard error about the average

[7] The average 'black' effect is a linear combination of the coefficients and its variance is then the corresponding quadratic form in the variance-covariance matrix. See also footnote 5.

effect is only one-quarter that of the unconstrained model, and thus the relatively small difference between black and white fertility is significant at the .004 level.

Finally, we conduct a test of whether the mean black effect (that is, the black log-odds minus the white log-odds) estimated with the unconstrained model is biased. The standard error of the difference in the mean black effect between unconstrained and the constrained models (0.247) is 0.159, or a *p*-value of .060. Taking the constrained model's mean 'black' effect as unbiased, we can say that the apparent upward bias in the unconstrained model's mean black effect, estimated by comparing the mean effect between the unconstrained and constrained models, is close to being statistically significant at the .05 level.

5. Summary and Discussion

We first described the statistical properties of the general constrained ML estimator. We then illustrated its application with a particular type of constraint function in which the weighted sum of all the covariate-specific associations (partial effects) of the regressors on the dependent variable is constrained to equal the overall population association of one or more of the regressors. We refer to those regressors whose bivariate or limited multivariate relationships with the dependent variable are constrained by population data as being "directly constrained." Our study investigated the improvements in the estimation of directly-constrained variables, and also the improvements in the estimation of other regressor variables that may be correlated with the directly-constrained variables, and thus "indirectly-constrained" by the population data.

We then showed with an empirical example that partial effects estimated from survey data may be re-estimated with a large reduction in variance by specifying population constraints on their overall association with the dependent variable. As an example application, we tested the minority-group status hypothesis of fertility on the marital fertility of black and white couples in which the wife is aged between 30 and 34 years old. According to the minority-group hypothesis, fertility will be lower for a minority group of otherwise equal economic status. The population data are age-, year- and race-specific marital fertility rates calculated from birth-registration data combined with census-based estimates of the married population by age and race. The survey data are from the Panel Study of Income Dynamics (PSID). These contribute additional information on pre-marital children, on duration of the current marriage, and on the economic condition of the family.

The standard errors about the directly-constrained coefficients were seen to be substantially, even drastically reduced under constrained MLE as compared to under unconstrained MLE. The indirectly-constrained coefficients, however, were changed negligibly in both point estimate and standard error about the estimate. We also used the comparison between constrained and unconstrained MLE to test for bias in the survey data, and found evidence of upward bias for white couples but not for black couples. Thus we have shown that the constrained MLE technique both provides a far more powerful statistical test of the minority-group hypothesis, and purges the test of a bias that would otherwise favor a finding in support of the minority-group hypothesis. We also found still extremely large reductions in the standard errors about the directly-constrained coefficients when further indirectly-constrained regressors were added. This indicates that directly constraining a regressor of interest can be expected to yield very large efficiency gains in a fully-specified regression model.

Substantively strong, but statistically weak support for the minority-group status hypothesis is found in our estimates from the unconstrained version of our equation. The

conclusion from the constrained version is that a statistically-significant but substantively-small minority-group effect is indicated. The difference in these conclusions from the point of view of the sociologist is potentially very large. As we noted above, earlier tests of the hypothesis for black versus white fertility have yielded mixed findings. These studies are not directly comparable to the present study, due among other things to the restriction of the present study to married women in their early 30s. Nevertheless, the results of the present study suggest that one factor contributing to the mixed findings in previous studies may be weaknesses in the empirical tests: either due to bias in the data for one or other of the black or white groups (as we found here for whites); or due to the large sampling error that arises in testing in relatively small subpopulations. It is indicative here that researchers more frequently turn to larger sample datasets such as the Census PUMS when analyzing racial and ethnic differences (e.g., Bean and Swicegood 1985), trading off sampling error for the problem of having fewer socio-economic variables available to specify a behavioral model. In the present study, we direct sociologists to an alternative approach that allows for the use of a smaller survey dataset and thus a more fully specified behavioral model, but using known population relationships to greatly increase the statistical power of the regression estimates.

As a caveat to the substantive finding here of a small minority-group status effect, the relatively simple set of socio-economic regressors used here may have resulted in marital fertility equations that are misspecified. If specification of the socio-economic variables were improved, for example to include more demographic information on previous births (total parity, parity within the present marriage, years since the previous birth) and/or to separate earnings of the husband and earnings or opportunity wages of the wife, the effect of the black coefficients may be reduced still further. In that case, we would be able to conclude against the operation of a substantial minority-group effect using a much more powerful statistical test than that with survey data only. The comparisons we made between the models estimated after adding more socio-economic regressors (the income-to-poverty ratio dummies) indicates that little reduction in statistical power would result from adding more regressors to the model.

Finally, we note that the statistical structure assumed for the population and survey data here will not always be realistic. In particular, we assumed that the population constraints were known exactly, in terms of both the associations of race and year with fertility (the marital fertility rates provided by NCHS) and the distribution of the regressor variables (estimated here from sample and not population data). Development and implementation of methods to account for uncertainty also in the constraints is a further topic that should be explored.

References

Bean, F.D., and G. Swicegood (1985) Mexican American Fertility Patterns Austin: University of Texas Press.

Bethlehem, J.G. (2002). Weighting nonresponse adjustments based on auxiliary information, in "Survey Nonresponse," Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. Editors, pp. 289-302, Wiley: New York.

Boyd, R.L. (1994) Educational mobility and the fertility of black and white women: A research note. Population Research and Policy Review 13(3):275-281.

Breslow, N.E. and Day, N.E. (1980). Statistical Methods in Cancer Research; Vol. 1, The Analysis of Case-Control Studies. IARC: Lyon.

Deming, W. E., and Stephan, F. F. (1942) On the Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Tables are Known. The Annals of Mathematical Statistics 11:427-424.

Elliott, M. R., and R. J. A. Little (2001) A Bayesian approach to combining information from a census, a coverage measurement survey and demographic analysis Journal of the American Statistical Association 95(450):351-362.

Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1998. An analysis of sample attrition in the Michigan Panel Study of Income Dynamics Journal of Human Resources 33:251-99.

Gelman, A. and Carlin, J. B. (2001) Poststratification and Weighting Adjustments, in "Survey Nonresponse," Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. Editors, pp. 289-302, Wiley: New York

Gelman, A. and Little, T. C. (1998) Improving Upon Probability Weighting for Household Size, Public Opinion Quarterly, 62, 398-404.

Goldscheider, C., and P.R. Uhlenberg (1969) Minority Group Status and Fertility American Journal of Sociology 74:361-372.

Handcock, M.S., S.M. Huovilainen, and M.S. Rendall. Combining Registration-System and Survey Data to Estimate Birth Probabilities Demography, 37(2):187-192.

Heitjan, D. F. (1990) Inference from Grouped Continuous Data; A Review (with discussion). Statistical Science 4:164-183.

Heitjan, D. F., and Rubin, D. B. (1989) Inference from Coarse Data via Multiple Imputation. Journal of the American Statistical Association 85(423):304-314.

Hellerstein, J., and G.W. Imbens (1999) Imposing moment restrictions from auxiliary data by weighting Review of Economics and Statistics 81(1):1-14.

Hill, M.S. 1992. The Panel Study of Income Dynamics: A User's Guide. Newbury Park, California: Sage.

Holt, D. and Smith T. F. M. (1979) Post-Stratification. Journal of the Royal Statistical Society, Series A. 142:33-46.

Imbens, G.W., Johnson, P. and Spady, R.H. (1998) Information Theoretic Approaches to Inference in Moment Condition Models, Econometrica 66: 333-59.

Imbens GW, Lancaster T. (1996) Efficient estimation and stratified sampling. Journal of Econometrics 74 (2): 289-318.

Imbens, G.W. and T. Lancaster (1994) Combining micro and macro data in microeconometric models. Review of Economic Studies 61: 655-680.

Ireland, C. T., and Kullback, S. (1968) Contingency tables with Given Marginals. Biometrika 55:179-188.

Johnson, N.E. (1988) The pace of births over the life course: implications for the minority-group status hypothesis Social Science Quarterly 69(1):95-107.

Kish, L. (1965) Survey Sampling New York: Wiley.

Little, R. J. A. (1991) Inference with Sample Weights, Journal of Official Statistics, 7, 405-424

Little, R. J. A. (1992) Regression with missing X's : A review Journal of the American Statistical Association 87(420):1227-1237.

Little, R. J. A. (1993) Post-stratification : A modeler's perspective Journal of the American Statistical Association 88(423):1001-1012.

Little, R. J. A. (1995) Missing data, in G. Arminger, C. C. Clogg, and M.E. Sobel (eds.) Handbook of Statistical Modeling for the Social Sciences New York: Chapman Hall.

Little, R. J. A. and Rubin D. B. (1987) Statistical Analysis of Missing Data. New York: John Wiley.

Little, R. J. A., and Wu, M. M. (1991) Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ Journal of the American Statistical Association 86(413):87-95.

Lohr, S. L. (1999) Sampling: Design and Analysis Pacific Grove, CA: Brooks-Cole.

Manski, C.F., and Lerman, S.R. (1977) Estimation of choice probabilities from choice based samples Econometrica 45 (8): 1977-1988.

National Center for Health Statistics (1999) Vital Statistics of the United States 1993. Volume 1 -- Natality. Hyattsville, MD: National Center for Health Statistics.

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

Qin, J., and J. Lawless (1994) Empirical likelihood and general estimating equations Annals of Statistics 22(1):300-325.

Rice J. A. (1995) Mathematical Statistics and Data Analysis Pacific Grove:Wadsworth.

Rubin, D. B. (1976) Inference and Missing Data. Biometrika 63:581-592.

Rubin, D. B. (1977) Formalizing Subjective Notions About The Effect on Non-respondents in Sample Surveys. Journal of the American Statistical Association 72(405):538-543.

SAS Institute (1997) SAS/OR Technical Report: The NLP Procedure Cary, NC: SAS Institute Inc.

Schafer J.L. (1997) Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Smith, D. P. (1992) Formal Demography New York: Plenum.

Smith, T. F. M. (1991) Post-Stratification. Statistician 40:315-323.

Table 1: Survey Covariate Distributions and Covariate-specific Birth Probabilities (Bivariate), Black and White Women Aged 30 to 35

	Proportion		
	Whites	Blacks	Birth Probability
Marital Duration			
1 year	0.034	0.043	0.013
2 years	0.040	0.056	0.112
3 years	0.045	0.057	0.191
4 years	0.052	0.066	0.191
5 years	0.059	0.069	0.177
6 years	0.063	0.064	0.163
7 years	0.062	0.068	0.167
8 years	0.070	0.071	0.144
9 years	0.075	0.076	0.127
10+ years	0.500	0.430	0.068
Pre-marital child			
none	0.926	0.806	0.116
one or more	0.074	0.194	0.035
Family Income-to-Poverty Ratio			
lowest third	0.311	0.503	0.074
middle third	0.336	0.330	0.100
top third	0.353	0.167	0.152
Sample N (person-year pairs t-1,t)	5,813	2,453	

Source: PSID 1984 to 1993

Note: Observations are of married couples in which the woman is aged 30 to 35 in the second year of the year pair.

Table 2: Unconstrained and Constrained Regression Estimates

Without income-to-poverty variable

	Unconstrained		Constrained		s.e.
	parameter	standard error	parameter	standard error (s.e.)	percent reduced
Intercept	-1.328 **	0.169	-1.564 **	0.115	31.8
Pre-marital child	-1.449 **	0.228	-1.441 **	0.227	0.1
year 1985	0.078	0.170	**0.027 ** **	**0.008**	95.2
year 1986	0.297	0.164	**0.056 ** **	**0.012**	92.4
year 1987	-0.119	0.178	**0.077 ** **	**0.010**	94.5
year 1988	0.115	0.170	**0.041 ** **	**0.008**	95.2
year 1989	0.028	0.168	**0.082 ** **	**0.009**	94.6
year 1990	-0.178	0.175	**0.114 ** **	**0.010**	94.1
year 1991	0.023	0.169	**0.040 ** **	**0.013**	92.2
year 1992	0.243	0.165	**0.053 ** **	**0.013**	92.2
year 1993	-0.098	0.172	**0.102 ** **	**0.011**	93.4
marriage duration 1 year	-2.864 **	0.541	-2.854 **	0.540	0.0
2 years	-0.622 **	0.212	-0.618 **	0.212	0.4
3 years	0.011	0.181	0.010	0.180	0.6
5 years	-0.106	0.171	-0.105	0.170	0.6
6 years	-0.220	0.172	-0.218	0.171	0.6
7 years	-0.195	0.171	-0.193	0.170	0.6
8 years	-0.345 *	0.171	-0.341 *	0.170	0.5
9 years	-0.541 **	0.173	-0.536 **	0.172	0.5
10 or more years	-1.277 **	0.138	-1.267 **	0.137	0.5
black	0.086	0.458	**-0.017**	**0.050**	89.0
black*pre-marital child	-0.023	0.617	-0.011	0.614	0.4
black*year 1985	-0.029	0.639	**-0.090 ** **	**0.027**	95.8
black*year 1986	-0.935	0.697	**-0.225 ** **	**0.025**	96.4
black*year 1987	0.109	0.618	**-0.338 ** **	**0.027**	95.6
black*year 1988	-1.242	0.774	**-0.338 ** **	**0.029**	96.2
black*year 1989	-0.645	0.636	**-0.414 ** **	**0.033**	94.8
black*year 1990	-0.687	0.692	**-0.291 ** **	**0.025**	96.3
black*year 1991	0.108	0.595	**-0.089 ** **	**0.019**	96.7
black*year 1992	-1.106	0.734	**-0.050 ** **	**0.022**	97.1
black*year 1993	-1.035	0.739	**-0.242 ** **	**0.023**	96.9
2nd tercile income-to-poverty	-	-	-	-	
3rd tercile income-to-poverty	-	-	-	-	
-2 log likelihood	-2673.3		-2703.6		

Notes: 1. Directly-constrained coefficients are shown in bold.

2. Statistically significant coefficients are indicated by * $p < .05$ and ** $p < .01$

Table 2: Unconstrained and Constrained Regression Estimates (continued)

With income-to-poverty variable

	Unconstrained			Constrained			s.e.
	parameter		standard error	parameter		standard error	percent reduced
Intercept	-1.628	**	0.185	-1.930	**	0.138	25.4
Pre-marital child	-1.319	**	0.229	-1.310	**	0.229	0.2
year 1985	0.081		0.170	**0.099**	****	**0.010**	94.3
year 1986	0.291		0.164	**0.120**	****	**0.009**	94.4
year 1987	-0.133		0.178	**0.132**	****	**0.007**	96.3
year 1988	0.095		0.171	**0.091**	****	**0.009**	94.6
year 1989	0.022		0.169	**0.145**	****	**0.008**	95.3
year 1990	-0.195		0.176	**0.168**	****	**0.008**	95.5
year 1991	0.019		0.169	**0.106**	****	**0.012**	93.0
year 1992	0.229		0.166	**0.106**	****	**0.012**	92.9
year 1993	-0.105		0.172	**0.165**	****	**0.008**	95.1
marriage duration 1 year	-2.833	**	0.541	-2.820	**	0.540	0.1
2 years	-0.611	**	0.213	-0.606	**	0.212	0.5
3 years	0.019		0.181	0.018		0.180	0.6
5 years	-0.104		0.172	-0.101		0.171	0.6
6 years	-0.222		0.172	-0.221		0.171	0.6
7 years	-0.190		0.172	-0.199		0.170	0.7
8 years	-0.316	*	0.172	-0.312		0.171	0.5
9 years	-0.503	**	0.173	-0.496	**	0.172	0.5
10 or more years	-1.173	**	0.140	-1.164	**	0.139	0.5
black	0.175		0.458	**0.141**	****	**0.054**	88.2
black*pre-marital child	-0.051		0.617	-0.047	**	0.615	0.4
black*year 1985	-0.056		0.639	**-0.186**	****	**0.030**	95.3
black*year 1986	-0.968		0.698	**-0.327**	****	**0.027**	96.1
black*year 1987	0.107		0.619	**-0.411**	****	**0.027**	95.6
black*year 1988	-1.221		0.774	**-0.385**	****	**0.029**	96.3
black*year 1989	-0.635		0.636	**-0.474**	****	**0.033**	94.9
black*year 1990	-0.721		0.692	**-0.393**	****	**0.028**	96.0
black*year 1991	0.089		0.596	**-0.129**	****	**0.022**	96.3
black*year 1992	-1.091		0.734	**-0.102**	****	**0.023**	96.9
black*year 1993	-1.003		0.738	**-0.281**	****	**0.025**	96.5
2nd tercile income-to-poverty	0.222	*	0.099	0.219	*	0.099	0.4
3rd tercile income-to-poverty	0.439	**	0.096	0.434	**	0.095	0.4
-2 log likelihood	-2662.4			-2693.7			

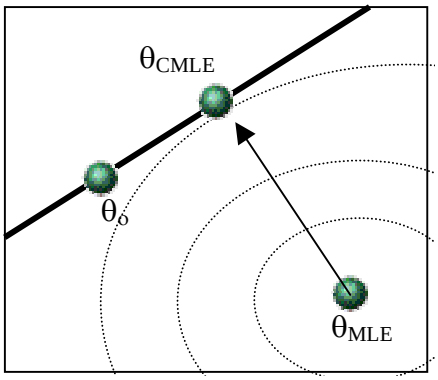


Figure 1: The relationship between estimates when representative sample data alone is used (MLE) and when it is augmented by exact population information (CMLE).

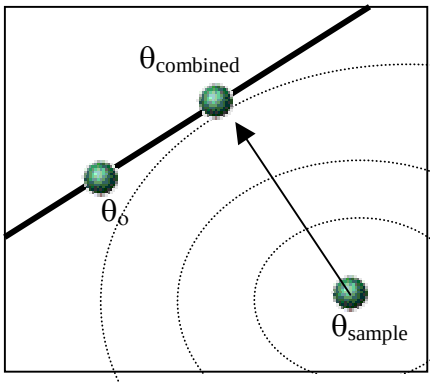


Figure 2: The relationship between the parameters of the unconstrained ("sample") and constrained ("combined") models and the population parameter when the sample survey is not representative.

Figure 3:

Survey (PSID) versus Population (NCHS) Marital Fertility Rates, Whites Aged 30 to 34

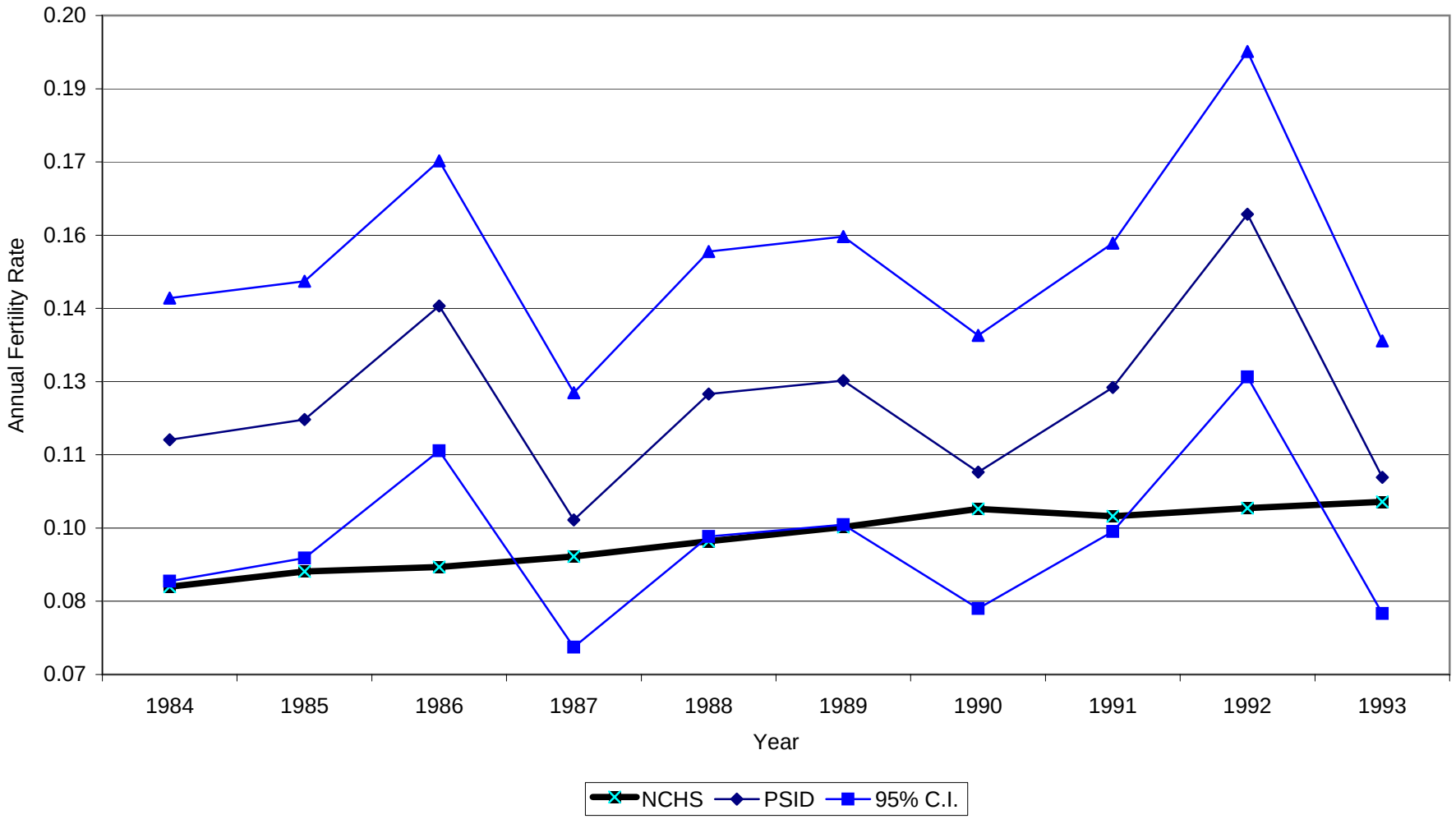


Figure 4:

Survey (PSID) versus Population (NCHS) Marital Fertility Rates, Blacks Aged 30 to 34

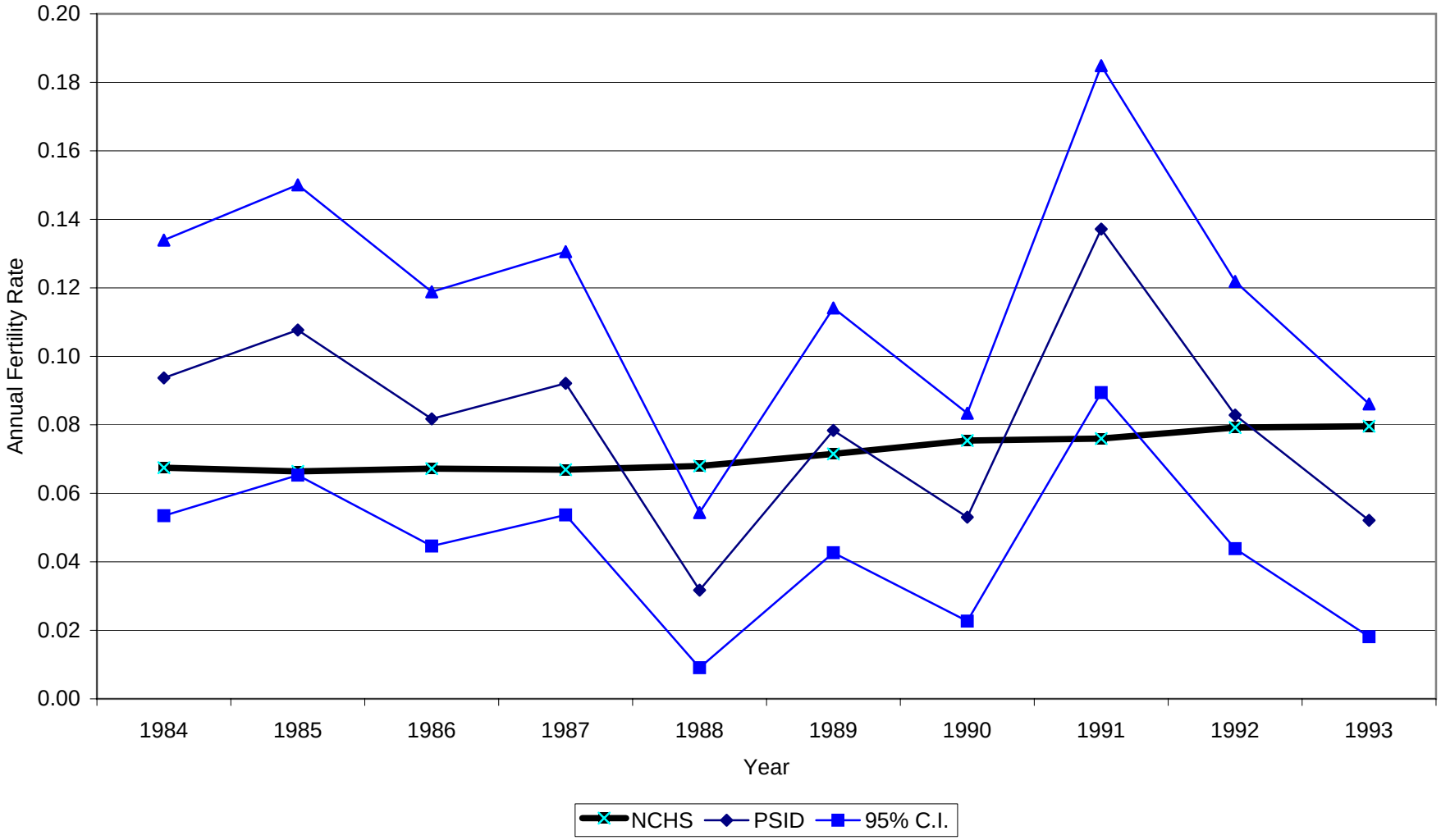
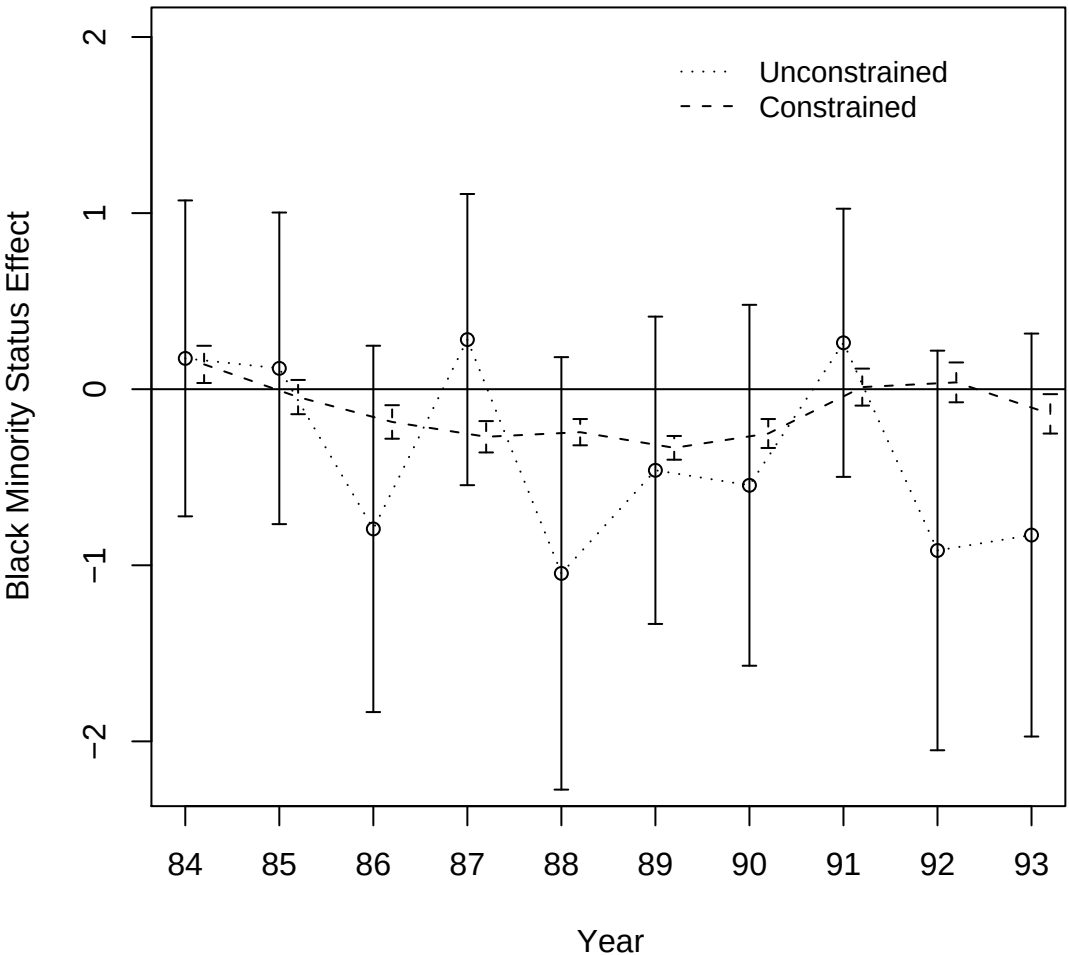
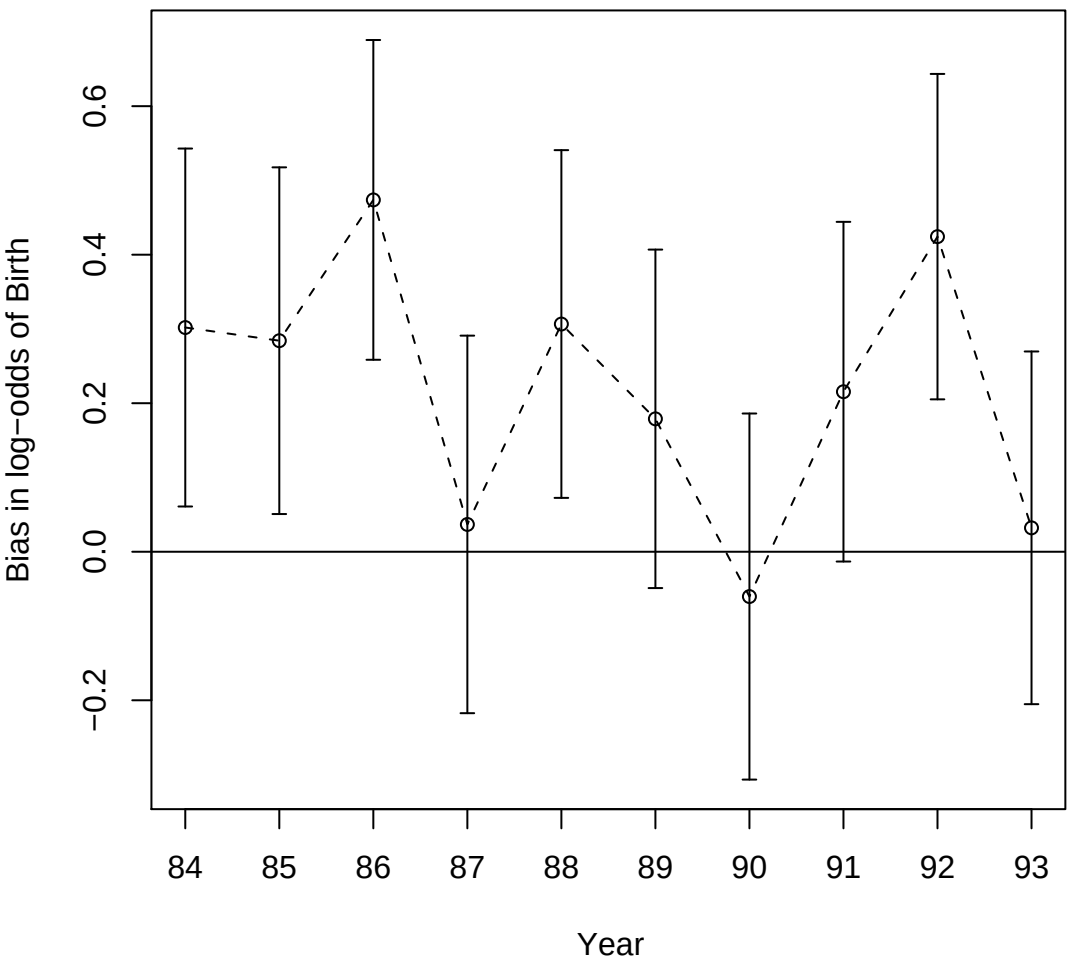


Figure 5: Estimates of the Black Minority−Status effect



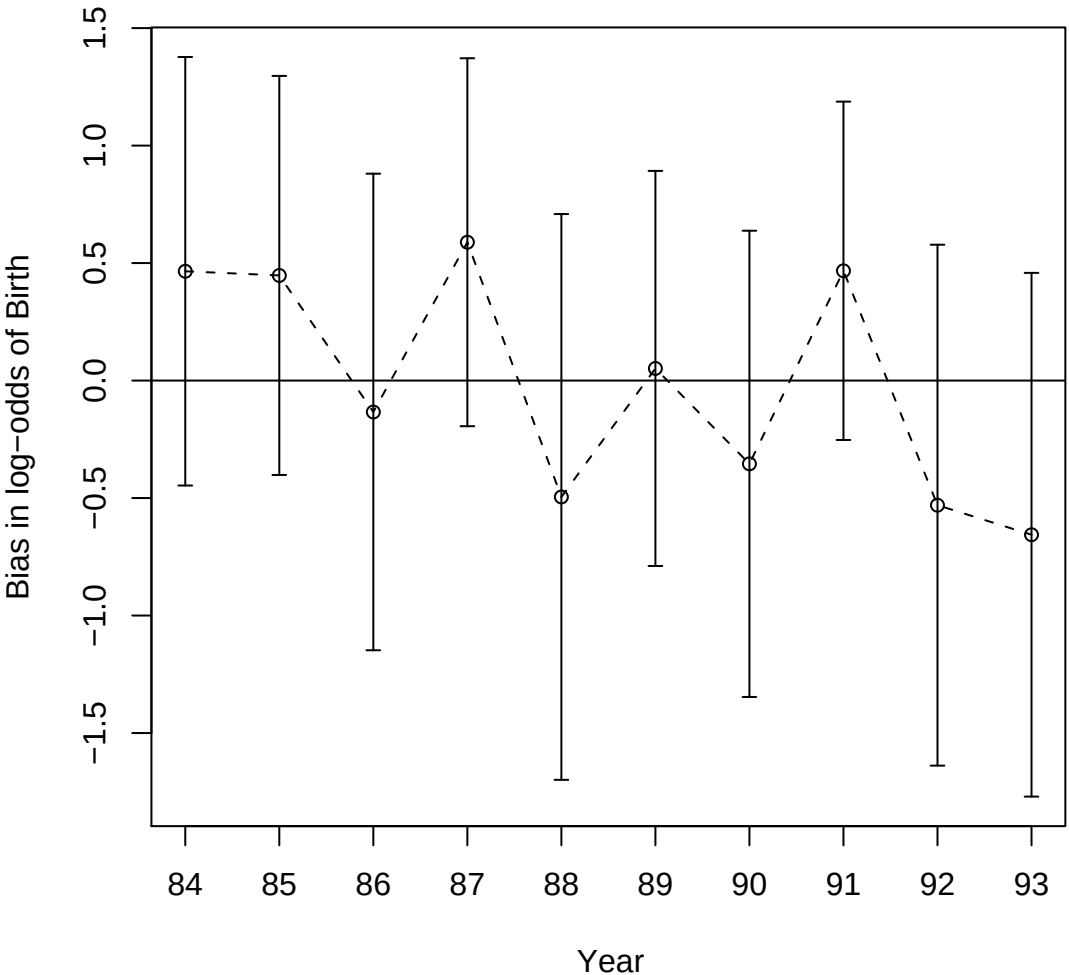
Note: Effect is estimated by the sum of the 'black' parameter estimate and the 'black−by−year' parameter estimate for that year

Figure 6: Estimates of the Bias of the PSID for white no pre−marriage child, married 4 years, and lowest income−to−poverty tercile



Note: Bias is estimated by the unconstrained log−odds minus the constrained log−odds

Figure 7: Estimates of the Bias of the PSID for black no pre−marriage child, married 4 years, and lowest income−to−poverty tercile



Note: Bias is estimated by the unconstrained log−odds minus the constrained log−odds