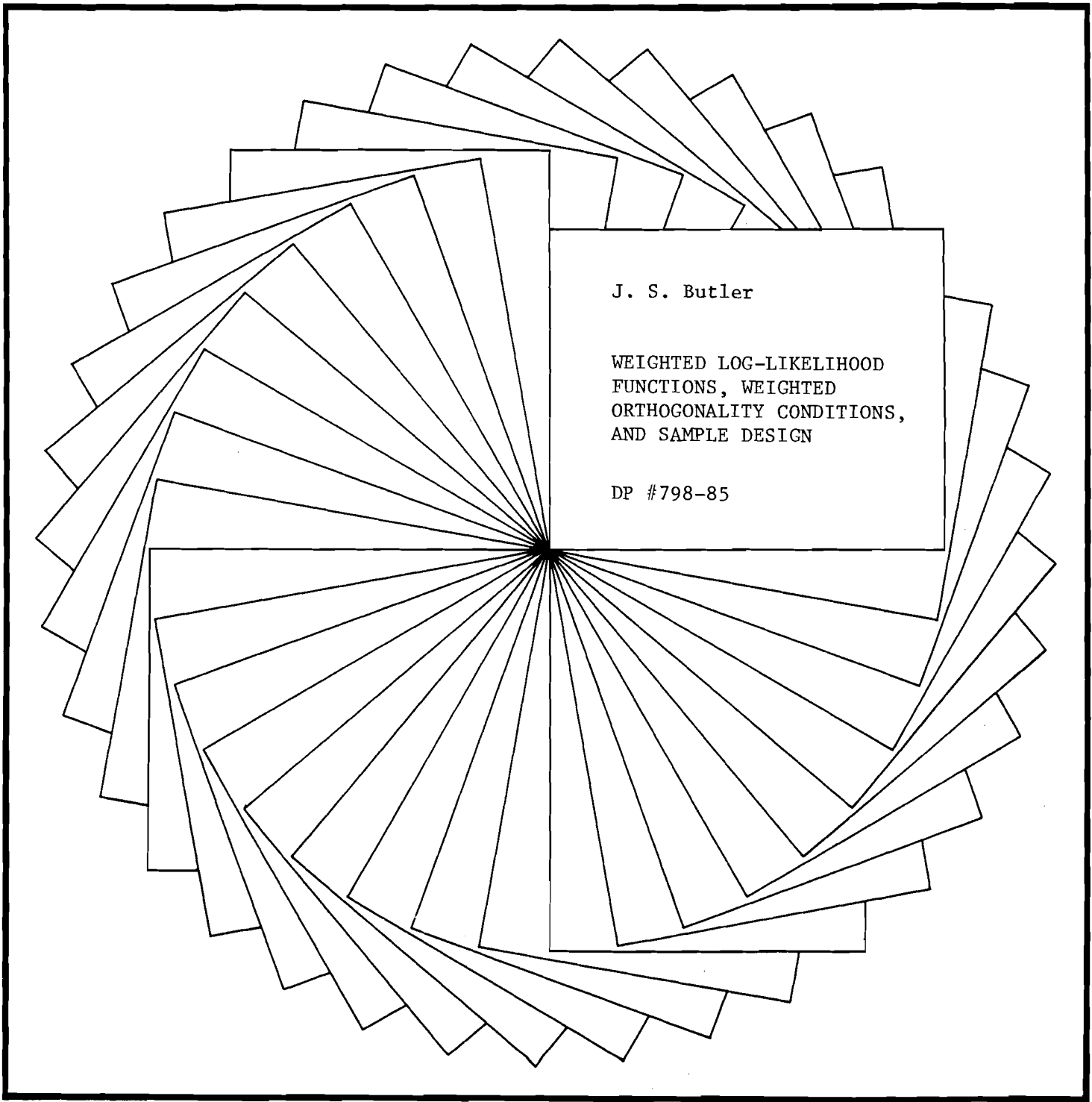


---

# IRP Discussion Papers

---



J. S. Butler

WEIGHTED LOG-LIKELIHOOD  
FUNCTIONS, WEIGHTED  
ORTHOGONALITY CONDITIONS,  
AND SAMPLE DESIGN

DP #798-85

Weighted Log-Likelihood Functions,  
Weighted Orthogonality Conditions, and Sample Design

J. S. Butler

Department of Economics  
Vanderbilt University

Revised, December 1985

This paper was presented as a seminar at the Institute for Research on Poverty, July 26, 1985. The author thanks Mathematica Policy Research for subsidizing this work. The opinions expressed in the paper are those of the author and do not necessarily reflect the views of the Institute or MPR. Comments from Randall Brown, William H. Greene, and Fran Seidita are appreciated. The author is responsible for any remaining errors.

## ABSTRACT

Virtually all survey data are weighted, such as those from the NIT experiments, the Health Interview Surveys, the National Ambulatory Care Survey, and the Current Population Survey. Weighted means are common in survey design, where the design effect is defined to be the ratio between the variance of the estimated mean from a stratified random sample and the variance from a simple random sample of the same size. Manski and Lerman (1977) derived the variance matrix for parameter estimates from weighted log likelihood functions. This paper shows that their results can be considerably simplified under certain conditions; the variance matrix is the design effect (a scalar) times the variance matrix from a simple random sample. In general, any estimate, a probit coefficient, a regression coefficient, a mean, etc., can be used as the basis of sample design.

The theory of sample design is well known (Cochran, 1977), and so is the theory of weighted likelihood functions (Coslett, 1981; Manski and Lerman, 1977). However, it is a fact that few researchers in economics use either sample design or weighted likelihood functions in applying maximum likelihood estimation. Many are unaware that variances change in well-known ways when data are weighted, and most cannot understand the necessary theory and programming to alter their computer packages to take account of the well-known results on weighted likelihood functions. Unfortunately, weighted data are ubiquitous in survey data, so the neglect of weighted likelihood functions is not merely an irritation, it is a general error. The purpose of this paper is to interpret some sample design for the case of weighted likelihood functions; to point out a potential simplification of the weighted likelihood functions under fairly extreme but not unknown conditions; to apply these results to the method of moments; and to suggest applications to sample design. None of this is mathematically difficult, but at present weighted likelihood functions seem totally impossible to most researchers. That need not be the case.

The first section rehashes the question of when weights are needed in regressions and other models. The second section derives the familiar design effect from weighted likelihood functions under a specific assumption. The third section relaxes the assumption somewhat. The fourth section discusses the application of these ideas to the method of moments. The fifth section presents an empirical example using these ideas. The sixth section suggests how these results could be used in designing surveys intended to produce more than sample means and cross-tabulations. A conclusion follows. An appendix reinterprets

standard results from sample design in terms of weighted log-likelihood functions.

#### Section 1: When Are Weights Required?

Almost all surveys are stratified, frequently on the basis of a variable of analytical interest such as income. The effects of stratification on the variances of estimated means is well known in survey analysis. (See, for example, Cochran, 1977.) Rarely has stratification been considered in the theory of likelihood functions, however. Hausman and Wise (1979) used a weighted maximum likelihood procedure. Manski and Lerman (1977) showed that the correct variance-covariance matrix from a weighted log likelihood function is not the usual inverted Hessian matrix or the inverted matrix of outer products of the first derivatives; instead, it is the outer product matrix pre- and post-multiplied by the inverted Hessian. (See Manski and Lerman, 1977, pp. 1984-5.) The asymptotic properties of the likelihood ratio test are then based on that product pre- and post-multiplied by the vector of differences between the true and estimated values of the parameters. A considerable simplification of those results is possible, requiring only a scalar correction to the customary and more easily computed variances and likelihood ratio tests calculated for maximum likelihood procedures.

Let the likelihood function be designated  $L$ , the log likelihood function  $L^*$ , the log likelihood of an observation on individual  $i$  be  $f_i$ , the first derivatives with respect to coefficients  $\underline{\theta}$  be  $L^*_{\underline{\theta}}$  or  $f_{\underline{\theta}}$ , and the second derivatives with respect to  $\underline{\theta}$  be  $L^*_{\underline{\theta}\underline{\theta}}$ , and  $f_{\underline{\theta}\underline{\theta}}$ .

Assume that a sample of data  $X$  is drawn with unequal weights, stratified in some way. Let  $y$  denote the outcome of interest. If the sample is stratified on exogenous variables only, and the object

of the analysis is to estimate endogenous variables given exogenous variables (and not the distribution of endogenous variables in the population), then the stratification may be ignored. If stratification is based on endogenous variables, e.g., choice-based sampling, then a weighted log likelihood function is required.

The difference between exogenous and endogenous stratification hinges on the disturbance in the equation. In essence, weights are needed when the weights are correlated with the disturbances in the statistical model. The regression setting illustrates this. Assume  $y = X\beta + \epsilon$  for endogenous  $y$ , exogenous  $X$ , and disturbance  $\epsilon$ . Stratifying on  $X$  alters the distribution of  $y$  but not  $\epsilon$ . Thus, statistics based ultimately on  $\epsilon$ , such as the variances of the estimated coefficients, are not affected by stratification on  $X$ . On the other hand, stratifying on  $y$  alters the distribution of  $\epsilon$  in the sample, including the variance. Note that if stratification is carried out on an exogenous variable,  $z$ , not included in the equation, then the distribution of disturbances is affected if  $z$  should have been included directly or interactively. A test of the difference of means can be written as a regression, but if the samples within the groups whose means are compared are themselves further stratified, then the distribution of the disturbance is affected if the further stratification involved variables which also should be in the regression. Note that omitted variable bias occurs only if the omitted variables are correlated with the dummy variable included in this implicit regression, but the distribution of  $\epsilon$  in the sample is changed as long as any stratifying variable is "omitted." The general points illustrated here hold for all statistical models; the question concerns the distribution of the disturbance.

To conclude: Weights are needed when the distribution of the disturbance must be repaired (see Hausman and Wise, 1976). Weights are therefore needed almost invariably when means are calculated from real surveys, because stratifying variables are found among the omitted variables, everything being omitted but a constant. In the test of independence of rows and columns in a cross-tabulation, no doubt stratifying variables are omitted, so weights are needed. In regression, weights are not needed. In choice-based samples, weights are needed.

## Section 2: The Familiar Design Effect; Weights Uncorrelated with Exogenous Variables

As in Manski and Lerman (1977), it is assumed that the probability mass or distribution function of the dependent variable in the sample and in the population are known. The ratio of the population probability distribution function (pdf) to that in the sample is the weight used in the likelihood function. We assume in this section that weights are stochastically independent of all exogenous and endogenous variables. That is true if the only regressor is a constant or if the regressors are sets of dummy variables uncorrelated with the weights. It is unlikely to be strictly true in general; it may or may not be far wrong in a specific case. One could test the hypothesis of independence of weights and exogenous variables, for example using a regression.

The weighted log likelihood function is  $\sum_{i=1}^N W_i f_i$ . The vector of first derivatives of  $f_i$  given  $\underline{\theta}^*$ , the maximum likelihood estimator, is independently and identically distributed with the usual mean (zero) and variance (the Hessian). The asymptotic expectation of the outer product of the first derivatives is the same as the asymptotic expectation of the Hessian;  $f$  is not a weighted log likelihood function.

$E(-f_{i\theta\theta'}) = E(f_{i\theta} f_{i\theta'})$  i.i.d. for all  $i$ . This usually leads to the assertion that

$$E(-L_{\theta\theta'}^*) = E(L_{\theta}^* L_{\theta'}^*), \quad (1)$$

used in deriving the properties of the maximum likelihood estimator and the likelihood ratio test. Here the result is different.

$$E(-L_{\theta\theta'}^*) = E\left(-\sum_{i=1}^N W_i f_{i\theta\theta'}\right) = \sum_{i=1}^N W_i E(-f_{i\theta\theta'}) \quad (2)$$

$$E(L_{\theta}^* L_{\theta'}^*) = E\left(\sum_{i=1}^N W_i^2 f_{i\theta} f_{i\theta'}\right) = \sum_{i=1}^N W_i^2 E(f_{i\theta} f_{i\theta'}) \quad (3)$$

The stochastic part of the model determines the nature and form of the likelihood function. Problems arise when weights are correlated with the disturbance. Suppose for argument that there is only one equation with one disturbance  $\epsilon$ . Person  $i$  draws a particular value of the disturbance and may or may not be in an oversampled stratum. If the same person were in the same population repeatedly having different  $\epsilon$ 's each time, and the  $\epsilon$ 's were observed only when person  $i$  was drawn into the sample, the observed  $\epsilon$ 's would trace out the unweighted likelihood function, which is not the true distribution of  $\epsilon$ 's, seen if person  $i$ 's  $\epsilon$  were observed every time. Weighting the values of  $f_i$  alters the sample distribution of  $\epsilon$  to reproduce the population distribution. However, the weights have another effect. They alter the form of the likelihood function as noted above, and they alter the form of the variance.

From (2) and (3) it follows that:



$$\frac{\sum W_i^2}{N} E(-L_{\underline{\theta}\underline{\theta}}^*) = E(L_{\underline{\theta}}^* L_{\underline{\theta}}^*). \quad (4)$$

If the weights were defined as  $W_i^*$  not summing to  $N$ , the scalar factor would be  $\Sigma(W_i^* \cdot N / \Sigma W_i^*)^2 / N = N^2 \Sigma W_i^{*2} / ((\Sigma W_i^*)^2 \cdot N) = N \Sigma W_i^{*2} / (\Sigma W_i^*)^2$ . The scalar factor exceeds one and thus  $E(-L_{\underline{\theta}\underline{\theta}}^*) < E(L_{\underline{\theta}}^* L_{\underline{\theta}}^*)$ .

With this result, one may proceed to evaluate the weighted variance derived by Manski and Lerman. They define  $\Omega = E(-L_{\underline{\theta}\underline{\theta}}^*)$  and  $\Delta = E(L_{\underline{\theta}}^* L_{\underline{\theta}}^*)$ . Let  $k = N / \sum_{i=1}^N W_i^2$ .

Then,

$$\Omega = \Delta k, \quad \Omega^{-1} = \Delta^{-1} / k, \quad \text{and} \quad \Omega^{-1} \Delta \Omega^{-1} = \Omega^{-1} / k = \Delta^{-1} / k^2.$$

Because  $k \leq 1$  for all positive weights  $W_i$ ,

$$\Omega^{-1} \Delta \Omega^{-1} \geq \Omega^{-1} \geq \Delta^{-1}. \quad (5)$$

More strongly,

$$|(\Omega^{-1} \Delta \Omega^{-1})_{ij}| \geq |\Omega^{ij}| \geq |\Delta^{ij}|, \quad \text{all } i \text{ and } j.$$

So, use of the conventional inverted Hessian biases all variances and covariances toward zero, and use of the outer product approximation does so even more. However,  $k$  is easily calculated in any sample, since it is  $k = (\Sigma W_i^2) / N \Sigma W_i^2$  for weights scaled to any sum.

The likelihood ratio statistic is biased upward. This follows from the fact that the variances are understated by using the conventional forms. The chi-square statistic is based on a Taylor expansion of the log likelihood function in the likelihood ratio,  $\lambda$  (see Theil, 1971, pp. 396-7)

$$2 \ln \lambda = 2(L^*(X, \underline{\theta}^*) - L^*(X, \underline{\hat{\theta}})), \quad (6)$$

where  $\hat{\theta}$  is the restricted maximum likelihood estimator. Then

$$2 \ln \lambda = (\underline{\theta}^* - \underline{\hat{\theta}})' \quad -\Sigma W_i^2 \frac{\partial^2 f_i(X, \theta^*)}{\partial \theta \partial \theta'} \quad (\underline{\theta}^* - \underline{\hat{\theta}}) \quad (7)$$

$$(2 \ln \lambda) = \frac{\Sigma W_i^2}{N} \sqrt{N} (\theta^* - \hat{\theta})' \quad - \frac{\partial^2 f_i(X, \theta^*)}{\partial \theta \partial \theta'} \quad (\theta^* - \hat{\theta}) \sqrt{N}. \quad (8)$$

The matrix in the middle of the right side of (8) has the same asymptotic distribution for all individuals, whether the sample is or is not weighted;  $f$  is not a weighted log likelihood function. Asymptotically, of the terms on the right side of (8), only  $\Sigma W_i^2$  depends on individuals' characteristics. The right side without the scalar  $\Sigma W_i^2$  constitutes the information matrix of a random sample size of  $N$ , and can be shown to follow a chi-square distribution with degrees of freedom equal to the number of restrictions involved. (See Silvey, 1975, pp. 113-4.) Thus, it follows from (8) that the likelihood ratio test is based on a number which must be multiplied by  $k$  to have the usual chi-square distribution. The size of the information matrix is overstated without the correction factor  $k$ ; this is equivalent to the statement that the covariance matrix is understated. Note that the correction factor is needed whether the difference between maximized likelihood functions is calculated using (6) or (8). The asymptotic distribution of that statistic is affected by weights.

None of the arguments depend on the distribution involved in setting up the likelihood function. The correction factor, therefore, does

not depend on the distribution involved, so long as the assumptions supporting maximum likelihood estimation hold.

Two standard examples of maximum likelihood estimation which fit the assumptions are presented in an appendix to illustrate these points: estimating the mean of a normal distribution with a weighted sample and testing the hypothesis of independence in a cross-classification using the likelihood ratio chi-square.

### Section 3. Weights Correlated with Endogenous Variables

Now  $E(f_{i\theta} f_{i\theta'})$  and  $E(f_{i\theta\theta'})$  are not constant across individuals because the weights are correlated with the explanatory variables. If the weights depended on the mode of travel chosen, as in Manski and Lerman, the same model that explains the mode chosen would explain the weights as well.

Assume that the distribution of data in  $X$  converges to an asymptotic distribution, which can depend upon the sample design, as the sample size grows. Then both the weighted outer product of first derivatives and the weighted second derivatives are functions of the data and parameters in the sense of Amemiya's (1973, p. 1002) statement of Jennrich's [1969] Theorem 1:

LEMMA 1 (Theorem 1 of Jennrich, 1969, p. 635): Let  $X$  be Euclidean space and  $\theta$  be a compact subset of a Euclidean space. If  $h$  is a bounded and continuous function on  $X \times \theta$ , and if  $\{G_T\}$  is a sequence of distribution functions on  $X$  which converge to a distribution function  $G$ , then

$$\int h(X, \theta) dG_T(X) \rightarrow \int h(X, \theta) dG(X)$$

uniformly for all  $\theta$  in  $\theta$ . For example

$$\int w(i)(f_{i\theta} f_{i\theta'}) dG_T(X) \rightarrow \int w(i)(f_{i\theta} f_{i\theta'}) dG(X).$$

The implication is that the weighted function follows a single distribution asymptotically. However, the weight is part of the function under the integral and cannot be separated if it is a function of the data  $X$ .

Note that  $E(f_{i\theta} f_{i\theta'})$  is not made a function of  $X$  by weighting which is based on values of explanatory variables. The values of  $f_{i\theta} f_{i\theta'}$  still constitute a random sample. So, even though the weights normally are functions of the data, frequently linear, the  $E(f_{i\theta} f_{i\theta'})$  are constant within categories of the outcome. Only the observed distribution of  $X$  is affected. Choice-based sampling requires weights based on the outcomes. However, within discrete categories of the outcome, either ranges of income in a weighted sample design from an N.I.T. experiment or discrete choices of mode of travel or participation versus nonparticipation, over persons  $i$  making choice  $j$ , or being in the range  $j$ ,

$$\frac{\sum W_i^2}{N_j} E(-L_{\theta\theta}^*) = E(L_{\theta\theta}^* L_{\theta\theta}^*).$$

Consequently, in Manski and Lerman's variance formula,  $\Omega^{-1} \Delta \Omega^{-1}$ , although a single scale factor cannot be derived, the calculation of second derivatives can be avoided. If the choices  $j=1$  to  $C$ , then

$$\Omega = \sum_{j=1}^C E(-L_{j\theta\theta}^*) = \sum_{j=1}^C \frac{N_j}{\sum W_{ij}^2} E(L_{j\theta} L_{j\theta'}) = \Delta.$$

$\Omega$  must be inverted, but it can be estimated by deflating each individual's outer product by a design effect calculated within the appropriate choice category. Thus,

$$\Omega = \sum_{j=1}^C A_j = \sum_{j=1}^C d_j B_j$$

$$\Delta = \sum_{i=1}^C B_j$$

and  $\Omega^{-1} \Delta \Omega^{-1} = (\sum d_j B_j)^{-1} (\sum B_j) (\sum d_j B_j)^{-1}$

where the A's and B's are appropriate positive definite matrices. Both  $\Omega$  and  $\Delta$  can be accumulated in one pass through a set of data.

#### Section 4: The Method of Moments

In brief, the method of moments with weighted data works as follows. See Hansen (1982). Orthogonality conditions are constructed:

$$\sum_{j=1}^N w_j g_j(\underline{\theta}) = \underline{0}_N(\underline{\theta}).$$

Note that  $g_j$  is a vector function, whose expected values are zero but not zero for any person  $j$ . They could be, e.g., products of exogenous variables and residuals. The function  $\underline{0}'_N \underline{0}_N$  is minimized by choosing  $\underline{\theta}$ . These estimates are used to estimate

$$D = \sum_{j=1}^N w_j \frac{\partial g_j(\underline{\theta})}{\partial \underline{\theta}}$$

and

$$V = \sum_{j=1}^N W_j^2 g_j(\underline{\theta}) g_j'(\underline{\theta}).$$

Then  $\underline{Q}'V^{-1}\underline{Q}$  is minimized. The  $\underline{\theta}$  thus obtained is distributed normally with a mean of the true parameter vector and variance  $D'V^{-1}D$ . In the method of moments, second derivatives are never needed. However, the weights are needed if the distribution of the disturbances is affected by the weighting (MOM is "fitting all models by OLS or GLS.") If the weights are not correlated with the orthogonality conditions or their first derivatives, then the same design effect holds--normalized weights,  $\Sigma w_j^2/N$  for  $N\Sigma w_j^2/(\Sigma w_j)^2$  in general. When the weights are correlated with the orthogonality conditions or their first derivatives, they can be analyzed analogously to the case of MLE--but weights should be taken into account.

#### Section 5: An Application to the Case of a Participation Equation

In this section, a study of the decision by elderly persons to participate in the Food Stamp Program is used to illustrate the ideas developed concerning weighted variances. First, the problem and the data are discussed briefly, then the results of the estimation are presented and discussed.

The Food Stamp Program is a welfare program under which many low-income households are eligible to receive stamps in dollar denominations tradeable for food in most stores. A persistent problem in the program from the point of view of proponents has been the low rate of acceptance of the stamps by eligibles, for which lack of information about the program and the stigma of welfare are possible explanations. A study

of elderly eligibles, the Food Stamp Cashout Project, was designed to examine factors associated with rejection of the stamps. (Cashout refers to payment in cash rather than stamps, a strategy for reducing the stigma.) That project is discussed in great detail in Butler, Ohls, and Posner (1985) and Blanchard et al. (1982).

Here, a probit model of participation in the program by 1,685 eligible persons is used to illustrate the effects of weighting. The sample design of the project entailed sampling based, in part, on participation, because the lists of participants reduce the cost of drawing a sample. That is the primary motivation mentioned by Manski and Lerman for such sampling. The explanatory variables include sociodemographic variables, age, race, sex, whether the eligible person lives alone, years of education, and an interaction between sex and living alone; geographical dummy variables standing for site (New York State, South Carolina, or Oregon) and whether cash was paid (a "demonstration" site); the potential bonus available and other gross income; frequency of getting out of the house (daily, weekly, or less often), distance to the Food Stamp office, and whether any bad experiences had ever been encountered there; and a measure of knowledge of nutrition: how many of the basic four food groups (breads, fruits and vegetables, meat and high-protein substitutes, and dairy products) are named by example in listing foods which should be eaten on a daily basis.

The data are summarized in Table 1. Maximum likelihood estimates are presented in Table 2, and method of moments estimates are presented in Table 3. The results are strikingly similar. Only one coefficient's sign ever changes, and that one, male-alone, has a t-value at most of 0.3. All magnitudes are extremely close to each other. Similarly,

Table 1

## Summary Statistics for Variables Used in the Participation Equation

	Mean	Standard Deviation	Maximum	Minimum
Constant	1.000	0.000	1	1
Age	73.855	6.176	99	65
Black	0.332	0.471	1	0
Male	0.296	0.457	18	0
Years of Education	7.269	3.699	18	0
Alone	0.816	0.387	1	0
NY Demonstration Site	0.091	0.287	1	0
NY Comparison Site	0.123	0.329	1	0
SC Demonstration Site	0.243	0.429	1	0
SC Comparison Site	0.228	0.420	1	0
OR Demonstration Site	0.148	0.355	1	0
OR Comparison Site	0.167	0.373	1	0
Food Stamp Bonus	35.253	24.372	183	10
Other Gross Income	301.540	116.500	1085	0
Out Daily	0.589	0.492	1	0
Out Weekly	0.277	0.447	1	0
Out Less Often	0.134	0.341	1	0
Male and Alone	0.144	0.351	1	0
Distance to Food Stamp Office	0.541	0.664	7	0
Bad Experiences at Food Stamp Office	0.052	0.221	1	0
Knowledge of Nutrition	2.072	1.111	4	0
Sampling Weight	3.171	3.789	19.803	0.518
Participation	0.675	0.468	1	0

Sample Size: 1685



Table 2

Maximum Likelihood Estimates of a Probit Model  
of Participation in the Food Stamp Program

	Unweighted		Weighted		Ratio of Standard Errors
	Coefficient	Standard Error	Coefficient	Standard Error	
Constant	4.684	0.398	5.229	0.703	1.769
Age - 65	-0.019	0.006	-0.023	0.008	1.459
Black	0.212	0.090	0.119	0.129	1.434
Male	-0.356	0.208	-0.174	0.348	1.670
Years of Education	-0.036	0.011	-0.051	0.016	1.390
Alone	-1.114	0.217	-1.380	0.357	1.645
NY Demonstration Site	0.560	0.156	0.542	0.196	1.258
NY Comparison Site	0.055	0.131	0.186	0.190	1.454
SC Demonstration Site	-0.251	0.122	-0.254	0.172	1.404
SC Comparison Site	-0.075	0.127	-0.073	0.174	1.365
OR Demonstration Site	0.193	0.128	0.150	0.210	1.643
FS Bonus (10's)	-0.157	0.020	-0.203	0.035	1.771
Gross Income (1000's)	-6.823	0.502	-7.681	1.137	2.265
Out Daily	-0.139	0.112	-0.232	0.153	1.362
Out Weekly	-0.068	0.120	-0.203	0.163	1.355
Male - Alone	0.063	0.233	0.050	0.376	1.615
Distance	-0.154	0.053	-0.187	0.070	1.322
Bad Experiences	-1.776	0.201	-1.876	0.235	1.170
Knowledge	0.033	0.032	0.054	0.046	1.442
Average Effect					1.515
Theoretical Effect					1.558
Sample Size = 1685					

Table 3

Method of Moments Estimates of a Probit Model  
of Participation in the Food Stamp Program

	Unweighted		Weighted		Ratio of Standard Errors
	Coefficient	Standard Error	Coefficient	Standard Error	
Constant	5.281	0.565	5.715	0.669	1.186
Age - 65	-0.019	0.006	-0.023	0.008	1.562
Black	0.193	0.089	0.103	0.132	1.479
Male	-0.366	0.252	-0.135	0.370	1.469
Years of Education	-0.034	0.011	-0.049	0.016	1.447
Alone	-1.309	0.288	-1.487	0.389	1.353
NY Demonstration Site	0.589	0.154	0.551	0.197	1.274
NY Comparison Site	0.100	0.140	0.227	0.196	1.397
SC Demonstration Site	-0.281	0.121	-0.290	0.176	1.465
SC Comparison Site	-0.106	0.124	-0.075	0.177	1.426
OR Demonstration Site	0.158	0.129	0.103	0.210	1.631
FS Bonus (10's)	-0.179	0.024	-0.225	0.031	1.310
Gross Income (1000's)	-8.031	0.939	-8.700	1.023	1.089
Out Daily	-0.147	0.109	-0.235	0.154	1.422
Out Weekly	-0.066	0.117	-0.187	0.166	1.411
Male - Alone	0.052	0.271	-0.001	0.397	1.464
Distance	-0.160	0.057	-0.195	0.068	1.199
Bad Experiences	-1.795	0.225	-1.917	0.252	1.120
Knowledge	0.035	0.031	0.048	0.047	1.469
Average Effect					1.378
Theoretical Effect					1.558
Sample Size = 1685					

the standard errors are very close in MLE and MOM estimates, but in both cases the standard errors are increased by taking into account the weights. The average effect, 1.378 under MOM and 1.515 under MLE, is close to the theoretical design effect for these data, 1.558. The t-values are overstated when weights are ignored.

#### Section 6: Implications for Sample Design

The ratio of the variance obtained using a particular sample design and sample size to the variance obtained using a simple random sample of the same size is the design effect. (See Cochran, 1977, p. 21, 98-101.) The sample design in a real survey normally involves trade-offs between the cost of the interviews and the variances of the statistical estimates obtained from the survey. In practice, the planning involves the assumption that sample means are to be estimated. Previous estimates of relevant means are obtained from other surveys, their variances are noted, and explicit trade-offs are calculated. However, the variances of regression coefficients, probit coefficients, or any other models can be used in the exercise, as long as an appropriate design effect is available.

Section 1 above shows that the usual design effect can apply to likelihood functions of all types, not just to estimated means from normally distributed populations. In planning a survey intended to generate maximum likelihood estimates, the design effect may be calculated as always and applied to any MLE--probit coefficients, t-values from regressions, chi-square statistics from cross-tabulations, etc.

Even though the design effect may vary for different parts of a population, except under restrictive conditions, it is possible that the design effect within subgroups of a population--bus riders and

auto drivers, e.g., or participants and non-participants in a welfare program--might not be too different, whereupon the same design effect can be used in planning a survey. If there are major differences among subpopulations, then a weighted average of them, weighted by variances or variance-covariance matrices, should be used in planning whenever maximum likelihood estimates or method of moments estimates of fancy models are intended, rather than making the assumption that the design effect for an estimated mean will be adequate. Modern econometrics does not depend on tests of differences of means, and neither should sample design.

#### Section 7. Summary and Conclusions

Weighted data are ubiquitous in social scientific research, but the use of weights is rare, despite well-known theoretical research in certain cases of weighting based on outcomes. Under certain conditions, encountered in traditional problems of sample design, a scalar factor can be applied to the inverse Hessian or outer-product first derivative matrix of maximum likelihood estimation to correct for the weighting. In other cases, more elaborate calculation is required. The method of moments is equally in need of weights in such cases; the calculations are little more complicated than those normally encountered in the MOM. An illustrative example is presented in which the variances are shown to increase with the application of choice-based weights, and the magnitudes are close to the theoretical design effect of sample design. Finally, in planning surveys, it is common to use the design effect to estimate the increase in variance of an estimated mean or proportion caused by stratification. This paper shows that any planned statistical activity can be used in designing

surveys. A probit coefficient, a regression coefficient, a sample mean, or any other statistical estimate with a variance can be used. Thus, surveys need not be planned as if only means and proportions were going to be estimated.

Appendix: Two Basic Examples of Weighted Likelihood Functions

Example 1. Estimating the Mean and Variance from a Normal Distribution

The following shows the derivation of the MLE of a mean and variance of a normal distribution from a weighted sample. The question arises why a weighted log-likelihood function would be used, since the minimum variance unbiased estimator of the mean of  $y$  is unweighted if  $y_t = \mu + \varepsilon_t$ ,  $\varepsilon_t$  i.i.d.  $N(0, \sigma^2)$ . In practice the assumption made is, rather, that  $y_t \sim N(\mu_t, \sigma^2)$  and that  $\mu_t = \underline{X}_t' \underline{\beta}$ . That is, a multivariate regression is implicit, possibly derived from a multivariate normal distribution of  $y_t$  and  $\underline{X}_t$ . Then weights are needed to generate an unbiased estimator of the mean of  $y$ :

$$\begin{aligned} E(y_t) &= E_{\underline{X}_t} (E(y_t | \underline{X}_t)) \\ &= E_{\underline{X}_t} (\underline{X}_t' \underline{\beta}) \\ &= \underline{\mu_X}' \underline{\beta}. \end{aligned}$$

Weights are needed whenever any form of stratification occurs. Whether or not the implicit regression is carried out, the weights applied to the sample of  $y$ 's produce an unbiased estimator.

The log likelihood function is  $L^*$ .

$$L^* = \sum_{i=1}^N W_i \ln \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(X_i - \mu)^2}{2\sigma^2} \quad (9)$$

$$L^* = - \frac{\sum W_i}{2} \ln(2\pi) - \frac{\sum W_i}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (W_i (X_i - \mu)^2) \quad (10)$$

The derivatives of  $L^*$  with respect to  $\mu$  and  $\sigma^2$  are

$$L_{\mu}^* = \frac{1}{2\sigma^2} (2\Sigma(W_i(X_i-\mu))) = 0 \rightarrow \mu = \Sigma W_i X_i / \Sigma W_i \quad (11)$$

$$L_{\sigma^2}^* = -\frac{\Sigma W_i}{2\sigma^2} + \frac{1}{2\sigma^4} \Sigma(W_i(X_i-\mu)^2) = 0 \rightarrow \sigma^2 = \Sigma(W_i(X_i-\mu)^2) / \Sigma W_i \quad (12)$$

$$L_{\mu\mu}^* = -(\Sigma W_i) / \sigma^2 \quad (13)$$

$$L_{\mu\sigma^2}^* = 0 \quad (14)$$

$$L_{\sigma^2\sigma^2}^* = +\frac{\Sigma W_i}{2\sigma^4} - \frac{2}{2\sigma^6} \Sigma(W_i(X_i-\mu)^2) = -\frac{\Sigma W_i}{2\sigma^2} \quad (15)$$

The matrices of second derivatives  $\Omega$  and of the outer products of first derivatives  $\Delta$  are as follows:

$$\Omega = \begin{vmatrix} -\frac{\Sigma W_i}{\sigma^2} & 0 \\ 0 & -\frac{\Sigma W_i}{2\sigma^4} \end{vmatrix} \quad -\Omega^{-1} = \begin{vmatrix} \frac{\sigma^2}{\Sigma W_i} & 0 \\ 0 & \frac{2\sigma^4}{\Sigma W_i} \end{vmatrix} \quad (16)$$

$$\Delta = E \begin{vmatrix} \frac{1}{\sigma^2} \Sigma W_i (X_i - \mu) \\ -\frac{\Sigma W_i}{2\sigma^2} + \frac{1}{2\sigma^4} \Sigma W_i (X_i - \mu)^2 \end{vmatrix} \begin{vmatrix} \frac{1}{\sigma^2} \Sigma W_i (X_i - \mu) & -\frac{\Sigma W_i}{2\sigma^2} + \frac{1}{2\sigma^4} \Sigma W_i (X_i - \mu)^2 \end{vmatrix} \quad (17)$$

$$\Delta_{11} = E \left[ \frac{1}{\sigma^4} (\Sigma W_i (X_i - \mu))^2 \right] = E \left[ \frac{1}{\sigma^4} \Sigma W_i^2 (X_i - \mu)^2 \right] = \frac{1}{\sigma^4} \Sigma W_i^2 \sigma^2 \quad (18)$$

$$\Delta_{12} = E \left[ -\frac{\sum W_j}{2\sigma^4} (\sum W_i (X_i - \mu)) + \frac{1}{2\sigma^6} (\sum W_i (X_i - \mu)) (\sum W_i (X_i - \mu)^2) \right] = E \left[ \frac{1}{2\sigma^6} \sum W_i^3 (X_i - \mu)^3 \right] = 0 \quad (19)$$

$$\Delta_{21} = E \left[ \frac{(\sum W_i)^2}{4\sigma^4} - \frac{2\sum W_i}{4\sigma^6} \sum W_i (X_i - \mu)^2 + \frac{1}{4\sigma^8} (\sum W_i (X_i - \mu)^2)^2 \right] \quad (20)$$

$$\Delta_{22} = \left[ \frac{(\sum W_i)^2}{4\sigma^4} - \frac{2(\sum W_i)^2 \sigma^2}{4\sigma^6} + \frac{\sum_j \sum_{i \neq j} W_j W_i \sigma^4 + \sum_i W_i^2 (3\sigma^4)}{4\sigma^8} \right] = \frac{2\sigma^4 \sum W_i^2}{4\sigma^8} = \frac{\sum W_i^2}{2\sigma^4} \quad (21)$$

From Manski and Lerman (1977, pp. 1984-5), the variance of the MLE of  $\mu$  and  $\sigma^2$  is (22).

$$\Omega^{-1} \Delta \Omega^{-1} = \begin{vmatrix} \frac{\sigma^2}{\sum W_i} & 0 \\ 0 & \frac{2\sigma^4}{\sum W_i} \end{vmatrix} \begin{vmatrix} \frac{\sum W_i^2}{\sigma^2} & 0 \\ 0 & \frac{\sum W_i^2}{2\sigma^4} \end{vmatrix} \begin{vmatrix} \frac{\sigma^2}{\sum W_i} & 0 \\ 0 & \frac{2\sigma^4}{\sum W_i} \end{vmatrix} = \begin{vmatrix} \sigma^2 \cdot \frac{\sum W_i^2}{(\sum W_i)^2} & 0 \\ 0 & 2\sigma^4 \cdot \frac{\sum W_i^2}{(\sum W_i)^2} \end{vmatrix} \quad (22)$$

where  $\sum W_i = N$ .

The correction factor  $k$  is  $N/\sum W_i^2$  in terms of normalized weights ( $\sum W_i = N$ ). For the variance of the estimator of  $\mu$ , the relevant values are these:  $\Omega^{-1}$  has  $\sigma^2/N$ ;  $\Delta^{-1}$  has  $\sigma^2/\sum W_i^2$ , and the correct matrix has  $\sigma^2 \sum W_i^2 / (\sum W_i)^2$ . The required equations hold:

$$\Omega^{-1} \Delta \Omega^{-1} = \Omega^{-1} / k \quad \frac{\sigma^2}{N} \div \frac{N}{\sum W_i^2} = \frac{\sigma^2 \sum W_i^2}{N^2} \quad (23)$$

$$\Omega^{-1} \Delta \Omega^{-1} = \Delta^{-1} / k^2 \quad \frac{\sigma^2}{\sum W_i^2} \div \frac{N^2}{(\sum W_i^2)^2} = \frac{\sigma^2 \sum W_i^2}{N^2} \quad (24)$$

$$\Omega^{-1} = \Delta^{-1} / k \quad \frac{\sigma^2}{\sum W_i^2} \div \frac{N}{\sum W_i^2} = \frac{\sigma^2}{N} \quad (25)$$



The results follow for the variances of the estimated  $\sigma^2$  with  $2\sigma^4$  substituted for  $\sigma^2$  in (23), (24), and (25).

Example 2. Likelihood Ratio  $\chi^2$  for a Test of the Independence of the Rows and Columns in a Table

The log likelihood function for multinomial sampling of an  $R \times C$  table with probabilities  $P_{ij}$ ,  $i=1$  to  $R$ ,  $j=1$  to  $C$ , is

$$L^* = \ln(n_{11}, n_{12}, \dots, n_{RC}) + \sum_{i=1}^R \sum_{j=1}^C n_{ij} W_{ij} \ln p_{ij}. \quad (26)$$

The number of observations in the sample in cell  $(i,j)$  is  $n_{ij}$ , and  $n$  is the sample size.  $W_{ij}$  is the average weight attached to observations in cell  $(i,j)$ . The weights serve the function of adjusting observed frequencies by known sampling rates in order to eliminate biases in relative proportions which would otherwise result. Assuming the weights to be normalized so that they sum to  $n$ , and letting  $W_{ij\ell}$  refer to an individual  $\ell$  in cell  $(i,j)$ , the concentrated log likelihood function is

$$L^* = \sum_{i=1}^R \sum_{j=1}^C \ln p_{ij} \sum_{\ell=1}^{n_{ij}} W_{ij\ell}. \quad (27)$$

This is maximized with respect to the  $R$  times  $C$   $p$ 's first with only the constraint  $1 = \sum_i \sum_j p_{ij}$ , then with the additional constraints that  $p_{ij} = p_{i+} p_{+j}$ , where the "+" denotes addition over a column or row, the constraint implies that the rows and columns are independently distributed. In the first or "unrestricted" case, the estimated  $p$ 's are given by

$$p_{ij} = \frac{\sum_{\ell=1}^{n_{ij}} W_{ij\ell}}{n}.$$

The maximized value of  $L^*$  is

$$\sum_{i=1}^R \sum_{j=1}^C \left( \sum_{\ell=1}^{n_{ij}} W_{ij\ell} \ln \frac{\sum_{\ell=1}^{n_{ij}} W_{ij\ell}}{n} \right) - n \ln n. \quad (28)$$

Under the "restricted" alternative hypothesis

$$p_{i+} = \frac{\sum_{j=1}^C \sum_{\ell=1}^{n_{ij}} W_{ij\ell}}{n}, \quad (29)$$

$$p_{+j} = \frac{\sum_{i=1}^R \sum_{\ell=1}^{n_{ij}} W_{ij\ell}}{n}, \quad (30)$$

and the maximized value of  $L^*$  is

$$\begin{aligned} & \sum_{i=1}^R \sum_{j=1}^C \sum_{\ell=1}^{n_{ij}} W_{ij\ell} \ln \frac{\sum_{j=1}^C \sum_{\ell=1}^{n_{ij}} W_{ij\ell}}{n} - n \ln n \\ & + \sum_{j=1}^C \sum_{i=1}^R \sum_{\ell=1}^{n_{ij}} W_{ij\ell} \ln \frac{\sum_{i=1}^R \sum_{\ell=1}^{n_{ij}} W_{ij\ell}}{n} - n \ln n. \end{aligned} \quad (31)$$

Twice the log of the likelihood ratio for this problem is (28) minus (31), which can be written as

$$2 \sum_{i=1}^R \sum_{j=1}^C \sum_{\ell=1}^{n_{ij}} W_{ij\ell} \ln \frac{\sum_{\ell=1}^{n_{ij}} W_{ij\ell}}{\left( \sum_{i=1}^R \sum_{\ell=1}^{n_{ij}} W_{ij\ell} \right) \cdot \left( \sum_{j=1}^C \sum_{\ell=1}^{n_{ij}} W_{ij\ell} \right)}. \quad (32)$$

This is the sum of observed frequencies times the log of the ratio of observed frequencies to expected frequencies under the restricted hypothesis, that is, the likelihood ratio chi-square. (The expected

frequency is the row total times the column total divided by  $n$ .) Bishop, Fienberg, and Holland (1975, pp. 513-8, 525-6) show that the likelihood ratio statistic follows a chi-square distribution with degrees of freedom equal to the sum of the eigenvalues of

$$D_{\pi}^{-0.5} \Sigma D_{\pi}^{-0.5}$$

in which  $D_{\pi}$  is a diagonal matrix with the vector of probabilities  $\underline{\pi}$  (estimated by  $\underline{p}$ ) on the main diagonal, and  $\Sigma$  is the variance-covariance matrix of the  $p$ 's. The eigenvalues are either zero or one in this case. The number of degrees of freedom is the difference between the number of unconstrained parameters in the constrained and unconstrained models, here  $(R-1) \times (C-1)$ .

With weighted data the variance matrix is multiplied by the design effect. This makes the eigenvalues equal to the design effect or zero. Thus the number of degrees of freedom in the chi-square distribution is multiplied by the design effect. A chi-square distribution must have an integral number of degrees of freedom, but a gamma distribution need not, in general. Dividing the likelihood ratio statistic by the design effect yields a statistic distributed as a chi-square with  $(R-1) \times (C-1)$  degrees of freedom.

Note that if stratification is based solely on variables which define the marginals of the table, then the weights may be ignored if there is no interaction effect. In this case, the chi-square statistic is asymptotically zero. If an interaction effect is present, then not only the marginal variable but also the interaction is stratified, and the chi-square statistic ignores that interaction, inasmuch as the test assumes its absence. Thus the test is biased toward rejecting

independence. The comparable situation in regression occurs if an interaction is omitted, the interaction involving a stratifying variable. In this case the exogeneity exception does not apply.