THE EFFECTS OF STRATIFICATION IN NONLINEAR SUPERPOPULATION MODELS

James Bethel, Westat, Inc.
1650 Research Boulevard, Rockville, Maryland 20850

1. Introduction

It is well known that stratification reduces the variance of direct expansion estimators by a factor of $1-\rho^2$ when the survey and stratification variables have a correlation ρ and are linearly related with homoscedastic regression errors. This result follows from the work of Cochran (1977) and Anderson, Kish and Cornell (1980), among others. In prior work, the author generalized this result to the case where the regression errors are heteroscedastic (Bethel 1988). In this paper we will extend it further to include nonlinear relationships between the survey and stratification variables. As it turns out, the asymptotic variances of optimally and proportionately allocated estimators do not depend on the function relating the survey and stratification variables. This result has some intuitive appeal: stratification should control the effects of $E(y \mid x)$ regardless of the functional form that it takes. In any case, the extension to nonlinear models allows us to consider an important class of models that describe lognormal survey variables.

We assume a finite population

$$\mathcal{P} = \{(x_1,y_1), (x_2,y_2), \ldots,(x_N,y_N)\}$$

of realizations of a random variable (Y,X) which satisfies the general regression model

$$(1) \quad Y = g(X) + v(X)\varepsilon.$$

We assume that X and ε are independent, with $E(\varepsilon) = 0$, $E(\varepsilon^2) = \sigma_\varepsilon^2 < \infty$, that X is bounded on a closed interval I, that g and v are continuous on I, and that $v(x) > 0$.

We will examine the asymptotic behavior of stratified estimators as L, the number of strata, increases. We do not specify the type of stratification to be used, but rather we simply assume that the stratum boundaries, $a_0 \le a_1 \le \cdots \le a_L$, satisfy

$$(2) \quad \max_{1\le h\le L} \{a_h - a_{h-1}\} \to 0 \quad \text{as} \quad L \to \infty.$$

When X is bounded, it can be shown that most common stratification strategies satisfy this condition.

The main results are that stratified estimators under proportionate allocation have

$$\mathrm{Var}\left(\left(\hat{\bar{Y}}_{PROP} - \bar{Y}\right) \mid \mathbf{X}\right) \to \sigma_\varepsilon^2 E_D\left(v(X_\alpha)^2\right)\left(\frac{1}{n} - \frac{1}{N}\right)$$

while optimally allocated estimators have

$$\mathrm{Var}\left(\left(\hat{\bar{Y}}_{OPT} - \bar{Y}\right) \mid \mathbf{X}\right)$$

$$\to \sigma_\varepsilon^2 \frac{1}{n}\left(E_D\left(v(X_\alpha)\right)\right)^2 - \sigma_\varepsilon^2 \frac{1}{N} E_D\, v(X_\alpha)^2$$

as $L \to \infty$. Empirical results given in Section 3 suggest that either of these stratified estimators is at least as efficient asymptotically as either ratio or regression estimators, particularly when g is nonlinear.

This paper is organized as follows. The next section presents the underlying asymptotic theory and derives the main results. Section 3 gives a numerical illustration and compares stratified, regression and ratio estimators. Section 4 discusses rates of convergence and Section 5 concludes with a discussion of our results.

2. Variances of Stratified Estimators

2.1 General Formulation

The population $\mathcal{P}$ is stratified as follows: Given a set of stratum boundaries $a_0, a_1, \ldots, a_L$, we label the pair (y_α,x_α) as (y_{hi},x_{hi}) – for some unique index i – whenever $a_{h-1} \le x_\alpha < a_h$ for $h < L$, or $a_{h-1} \le x_\alpha \le a_h$ when $h = L$. We assume that simple random sampling is used within strata. We will consider stratified estimators of the mean $\bar{Y}$:

$$(3) \quad \hat{\bar{Y}} = \sum_{h=1}^{L} W_h \bar{y}_h$$

where $\bar{y}_h$ and N_h are the mean and size of the h-th stratum, respectively, and $W_h = N_h/N$. As L increases, we must have $N \ge n \ge L$. Depending on the method of sample allocation, there may be conditions on n_h which introduce additional restrictions on n (and thus N). Beyond recognizing these implicit relations, however, we will make no specific assumptions about the rates at which n and N increase.

In what follows, it will be important to keep in mind that there are two sources of randomness: the first results from generating the pairs in $\mathcal{P}$ according to the model in (1) – which we will call model randomness – while the second results from the selection of a sample of n pairs from $\mathcal{P}$ according to the sample design – which we will call design randomness. We will denote expected value with respect to model randomness by $E_M(\cdot)$, and expected value with respect to design randomness by $E_D(\cdot)$. Notation for variances will use the same convention. $E_D(\cdot)$ and $V_D(\cdot)$, depending on the context, may represent the mean and variance under either stratified or simple random sampling. The mean with respect to design randomness within stratum h will be denoted as $E_D^{(h)}(\cdot)$.

For the most part, we will consider the conditional variance of $\hat{\bar{Y}} - \bar{Y}$ given the values of $X_1, X_2, \ldots, X_N$. Assuming unbiasedness,

$$\mathrm{Var}\left(\hat{\bar{Y}} - \bar{Y} \mid \mathbf{X}\right) = E_M\left(V_D\left(\hat{\bar{Y}} - \bar{Y}\right) \mid \mathbf{X}\right)$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_N)$.

2.2 Variance under Proportionate Allocation

First we note that

$$(5) \quad E_M\left(V_D\left(\hat{\bar{Y}} - \bar{Y}\right) \mid \mathbf{X}\right) = E_M\left(\sum_{h=1}^{L} W_h^2 S_h^2(y_{hi})\left(\frac{1}{n_h} - \frac{1}{N_h}\right) \mid \mathbf{X}\right),$$

where $S_h^2(\cdot)$ is the variance of the argument within stratum h under simple random sampling. Furthermore

$$E_M\left(S_h^2(y_{hi}) \mid \mathbf{X}\right)$$

$$= E_M\left(S_h^2(g(x_{hi})) + S_h^2(v(x_{hi})\varepsilon_{hi}) + 2S_h(g(x_{hi}),v(x_{hi})\varepsilon_{hi}) \mid \mathbf{X}\right)$$

$$= S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_M\left(\frac{1}{N_h}\sum_{i=1}^{N_h} v(X_{hi})^2 \mid \mathbf{X}\right)$$

$$= S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right),$$

where $S_h(\cdot\,,\cdot)$ is the stratum covariance of the arguments under simple random sampling.) Thus we have

$$(6)\quad E_M\left(V_D\left(\hat{\bar{Y}} - \bar{Y}\right)\mid X\right) = \sum_{h=1}^{L} W_h^2 S_h^2(g(X_{hi}))\left(\frac{1}{n_h} - \frac{1}{N_h}\right)$$

$$+ \sum_{h=1}^{L} W_h^2\,\sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right)\left(\frac{1}{n_h} - \frac{1}{N_h}\right).$$

From the uniform continuity theorem, g(x) is uniformly continuous on the domain of X, and, from equation (2),

$$\mid x_{hi} - x_{hj} \mid \;\le \underset{1\le h\le L}{\text{Max}}\,\{a_h - a_{h-1}\} \;=\; o(1)$$

Thus $S_h^2(g(x_{hi}))$ is o(1) as $L \to \infty$, from which it follows that the first term in (6) is also o(1). From (6) we have

$$(7)\quad E_M\left(V_D\left(\hat{\bar{Y}} - \bar{Y}\right)\mid X\right) = o(1)$$

$$+ \sigma_\varepsilon^2 \sum_{h=1}^{L} W_h^2 E_D^{(h)}\left(v(X_{hi})\right)^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right).$$

$$= o(1) + \sigma_\varepsilon^2 \sum_{h=1}^{L} W_h^2 E_D^{(h)}\left(v(X_{hi})^2\right)\frac{1}{n_h} \;-\; \sigma_\varepsilon^2 E_D\left(v(X_\alpha)^2\right)\frac{1}{N}.$$

Under proportionate allocation, $n_h = nW_h$, thus (7) becomes

$$(8)\quad E_M\left(V_D\left(\hat{\bar{Y}}_{PROP} - \bar{Y}\right)\mid X\right)$$

$$= o(1) + \sigma_\varepsilon^2 \sum_{h=1}^{L} W_h E_D^{(h)}\left(v(X_{hi})^2\right)\frac{1}{n} \;-\; \sigma_\varepsilon^2 E_D\left(v(X_\alpha)^2\right)\frac{1}{N}$$

$$\to \sigma_\varepsilon^2 E_D\left(v(X_\alpha)^2\right)\left(\frac{1}{n} - \frac{1}{N}\right)$$

giving the asymptotic variance under proportionate sampling.

2.3 Variance Under Optimum Allocation

As we saw above,

$$E_M\left(S_h^2(y_{hi})\mid X\right) = S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right).$$

Under optimum allocation,

$$n_h = n\,\frac{W_h\sqrt{S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right)}}{\sum_{h=1}^{L} W_h\sqrt{S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right)}}$$

thus, from (5), we have

$$(9)\quad E_M\left(V_D\left(\hat{\bar{Y}}_{OPT} - \bar{Y}\right)\mid X\right)$$

$$= E_M\left(\sum_{h=1}^{L} W_h^2 S_h^2(y_{hi})\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\mid X\right)$$

$$= \left(\sum_{h=1}^{L} W_h\sqrt{S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right)}\right)^2\frac{1}{n}$$

$$+ o(1) + \sigma_\varepsilon^2 E_D\left(v(X_\alpha)^2\right)\frac{1}{N}.$$

Let $m_h = (a_h + a_{h-1})/2$. Clearly

$$X_{hi} \xrightarrow{\mathcal{D}} m_h \qquad \text{as } L \to \infty$$

(by which we mean that $P\left(X \le t \mid a_{h-1} \le X \le a_h\right) \to 1$ if $X \ge m_h$ and 0 otherwise). It follows that

$$v(X_{hi})^2 \xrightarrow{\mathcal{D}} v(m_h)^2.$$

Since X is bounded, the sequence $\{X_{hi}\}$ is uniformly integrable, and thus

$$E_D^{(h)}\left(v(x_{hi})\right) \to v(m_h)$$

and

$$E_D^{(h)}\left(v(x_{hi})^2\right) \to v(m_h)^2.$$

It follows that

$$(10)\quad \sqrt{E_D^{(h)}\left(v(x_{hi})^2\right)} \to \sqrt{v(m_h)^2} = v(m_h).$$

Combining these results, we have

$$\sqrt{E_D^{(h)}\left(v(x_{hi})^2\right)} \;-\; E_D^{(h)}\left(v(x_{hi})\right) = o(1).$$

Since $S_h^2(g(x_{hi})) = o(1)$, formula (10) gives

$$(11)\quad \sqrt{S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right)} \;\to\; \sigma_\varepsilon E_D^{(h)}\left(v(x_{hi})\right).$$

Finally, applying the bounded convergence theorem and combining (9) - (11), we obtain

$$E_M\left(V_D\left(\hat{\bar{Y}} - \bar{Y}\right)\mid X\right)$$

$$= \left(\sum_{h=1}^{L} W_h\sqrt{S_h^2(g(X_{hi})) + \sigma_\varepsilon^2 E_D^{(h)}\left(v(X_{hi})^2\right)}\right)^2\frac{1}{n}$$

$$+ o(1) + \sigma_\varepsilon^2 E_D\left(v(X_\alpha)^2\right)\frac{1}{N}$$

$$\to \sigma_\varepsilon^2\frac{1}{n}\left(\sum_{h=1}^{L} W_h E_D^{(h)}\left(v(X_{hi})\right)\right)^2 \;-\; \sigma_\varepsilon^2\frac{1}{N}E_D\,v(X_\alpha)^2$$

$$= \sigma_\varepsilon^2\frac{1}{n}\left(E_D(v(X_\alpha))\right)^2 \;-\; \sigma_\varepsilon^2\frac{1}{N}E_D\,v(X_\alpha)^2$$

which gives the asymptotic variance under optimum allocation.

3. Application to Log Transportation Model

As an application, we consider the log transformation model. This model is given by

$$y = x^\alpha\varepsilon'$$

where x and ε' are independent lognormal random variables. Notice that ε' does not satisfy $E(\varepsilon') = 0$, but that is easily accomplished by setting $\varepsilon = \varepsilon' - E(\varepsilon')$. Then

$$y = x^\alpha E(\varepsilon') + x^\alpha\varepsilon$$

This goes beyond the linear model that is generally used to analyze sampling strategies, but it is included in the generalized model in (1).

This model can be parameterized by the mean and standard deviation of the distributions of $ln\ x$ and $ln\ \varepsilon$, together with the exponent α. However, since the mean of $ln\ x$ and $ln\ \varepsilon$ have no effect on the relative efficiencies of the estimators we are considering, we will take them to be zero. Thus we assume that

$$ln\ x \sim N(0,\tau_1^2) \text{ and } ln\ \varepsilon' \sim N(0,\tau_2^2).$$

Table 1 shows the asymptotic efficiencies of ratio, regression, and optimally and proportionately allocated stratified estimators, as compared with simple random sampling. The parameters τ_1, τ_2, and α are varied over the ranges of .50-1.00, .25-1.00, and .50-1.50, respectively. This table suggests several conclusions. First, although proportionately allocated estimators are always better than regression or ratio estimators, the improvement is small when τ_2 is large, and vice versa. Second, optimum allocation is significantly more efficient than simple random sampling (or any of the other methods) as either τ_1 or α increases. Finally, although the ratio estimator shows some significant savings over SRS for $\alpha = 1.50$, it performs worse for $\alpha = .50$.

Table 1: Efficiencies of stratified, regression, and ratio estimators*

$Var(ln\ x)$ τ_1	$Var(ln\ \varepsilon)$ τ_2		Model Exponent α 0.50	0.75	1.50
0.50	0.25	Optimum Allocation	2.06	3.49	13.46
		Proportionate Allocation	1.94	3.03	7.67
		Regression	1.88	2.98	6.23
		Ratio	0.88	2.33	3.58
	0.50	Optimum Allocation	1.29	1.68	4.41
		Proportionate Allocation	1.21	1.46	2.51
		Regression	1.21	1.46	2.39
		Ratio	0.95	1.37	2.00
	1.00	Optimum Allocation	1.10	1.24	2.19
		Proportionate Allocation	1.04	1.08	1.25
		Regression	1.03	1.08	1.24
		Ratio	0.99	1.06	1.20
0.75	0.25	Optimum Allocation	3.49	7.14	43.01
		Proportionate Allocation	3.03	5.21	12.13
		Regression	2.63	4.80	6.18
		Ratio	0.68	2.77	3.16
	0.50	Optimum Allocation	1.68	2.68	12.51
		Proportionate Allocation	1.46	1.95	3.53
		Regression	1.41	1.92	2.90
		Ratio	0.82	1.63	2.14
	1.00	Optimum Allocation	1.24	1.59	5.03
		Proportionate Allocation	1.08	1.16	1.42
		Regression	1.07	1.15	1.37
		Ratio	0.95	1.12	1.28
1.00	0.25	Optimum Allocation	5.69	13.46	141.09
		Proportionate Allocation	4.43	7.67	14.87
		Regression	3.01	6.11	4.45
		Ratio	0.47	2.46	2.36
	0.50	Optimum Allocation	2.28	4.41	39.37
		Proportionate Allocation	1.78	2.51	4.15
		Regression	1.61	2.38	2.71
		Ratio	0.61	1.70	1.88
	1.00	Optimum Allocation	1.45	2.19	14.43
		Proportionate Allocation	1.13	1.25	1.52
		Regression	1.11	1.24	1.40
		Ratio	0.86	1.16	1.27

*The values given are the efficiencies of the estimator as compared with direct expansion the estimator under simple random sampling. Thus

$$\text{Efficiency of Estimator} = \frac{\text{SRS Variance}}{\text{Variance of Estimator}}$$

4. Rates of Convergence

It seems unlikely that any rigorous results on rates of convergence can be obtained without some stronger assumption about g(x), such as, for example, that g' is bounded. For the special case where g is linear, arguments similar to those used here suggest that the convergence is at rate $O(L^{-2})$ (Cochran 1977, Bethel 1988).

Table 2 shows the efficiency of optimally and proportionately allocated estimates with the L = 2, 5, 10, 20, and ∞, under the log transformation model with $\tau_1 = 1.0$, $\tau_2 = 0.5$, and $\alpha = 0.75$. Stratum boundaries were defined using the Dalenius-Hodges (1959) technique. Analysis of the data in this table suggests that the rate of convergence in the log transformation model is about $O(L^{-1.55})$ under proportionate allocation and $O(L^{-1.75})$ under optimal allocation, at least for L > 2.

Table 2: Rates of Convergence in the Log Transformation Model*

Number of Strata	Efficiency V_{OPT}	V_{PROP}
2	2.12	1.54
5	3.61	2.17
10	4.14	2.39
20	4.33	2.47
∞	4.41	2.51

*Calculations assume $t_1 = 1.0$, $t_2 = 0.5$, and a = 0.75. Efficiency is defined as in Table 1. Dalenius-Hodges boundaries were used to define strata.

5. Discussion

It generally recognized that regression and ratio estimators should be stratified when the slope of the regression line varies between strata, as would be the case when g(x) is nonlinear (e.g., see Des Raj 1977). From this point of view, estimators would be expected to be superior for nonlinear models, although the fact that a proportionately allocated estimate might be two or three times as efficient as either a regression or ratio estimate (e.g., see Table 1 where $\alpha = 1.5$, $\tau_1 = 1.0$, $.25 \le \tau_2 \le .50$) is rather surprising.

Concerning the issue of "combined versus separate" regression and ratio estimators, presumably the latter would be preferable when g is nonlinear. However, following the arguments given in Section 2.3, $S_h(Y_{hi},X_{hi}) \to 0$ for large L, so that it is not clear whether separate regression or ratio estimators would improve on stratified, optimally allocated direct expansion estimators. When g(x) is linear, Wright (1983) shows that combined regression and ratio estimators asymptotically achieve the variance shown in the RHS of (9) under unequal probability sampling. Although it is uncertain what this implies for stratified sampling when g(x) is nonlinear, it seems likely that stratified, separate regression or ratio estimators would be as efficient as stratifed estimators under optimal allocation.

Generally speaking, optimal allocation is not a feasible strategy, since the required data on stratum variances are usually not available. On the other hand, proportionate allocation, regression estimation, and ratio estimation are feasible. The results derived here show that stratifed estimators using proportionate allocation are preferable to separate regression or ratio estimators. As noted above, it is not clear whether combined regression or ratio estimators

would be more efficient. The results on proportionate allocation also have implications for post-stratification, since post-stratified estimates behave much like stratified estimates under proportionate allocation. As a final note, we point out that the results derived here are asymptotic: either ratio or regression estimators might be more efficient for small sample sizes.

REFERENCES

Anderson, D.W., Kish, L., and Cornell, R.G. (1980). On stratification, grouping, and matching. *Scandinavian Journal of Statistics*, **7**, 61-66.

Bethel, J. (1988). Minimum variance estimation in stratified sampling. *Journal of the American Statistical Association*, to appear.

Cochran, W.G. (1977). *Sampling Techniques*, John Wiley & Sons, New York.

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, **54**, 88-101.

Des Raj (1977). *Sampling Theory*, McGraw-Hill, New York.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, **78**, 879-884.