MODELING INTERVIEWER VARIABILITY FOR DICHOTOMOUS VARIABLES

S. Lynne Stokes and Joe R. Hill, University of Texas at Austin

1. INTRODUCTION

Measurement of the correlated component of response error in sample surveys is important for three main reasons. First, if the correlation is due to the interviewers, as we will assume throughout this paper, a large correlation may indicate a problem with interviewer training on that item or with the design of the question itself. If such items can be identified, they might be improved.

Second, the presence of the correlation inflates the variance of the sample mean $\tilde{\mu}_p$, which is commonly used as the estimator of the population mean of the item responses. Let y_{ij} be the response of the j^{th} unit of interviewer i's assignment, $i = 1,..., k$; $j = 1, ..., n_i$. Writing $N = \Sigma n_i$, $\overline{n-1} = \Sigma n_i(n_i - 1)/N$, we have

$$\begin{aligned} Var(\tilde{\mu}_p) &= Var(\Sigma\Sigma\, y_{ij} / N) \\ &= [Var\,(y_{ij}) + (\overline{n-1})\, Cov(y_{ij},y_{ij'})]/ N \\ &= Var(y_{ij})\,[1 + (\overline{n-1})\,\rho]/N, \qquad (1.1) \end{aligned}$$

where $\rho = Cov(y_{ij},y_{ij'})/Var(y_{ij})$. If $(\overline{n-1})$ is moderately large, even a very small ρ can dramatically increase the variance of $\tilde{\mu}_p$.

Third, an estimate of the correlation ρ can be used to improve the estimate of the population mean μ_p. Letting $y_i = \sum_j y_{ij} / n_i$, then the best linear unbiassed estimator of μ_p based on the y_i's has optimal weights equal to the inverse of the variances of the y_i's, which depend on ρ. In particular, $\hat{\mu}_p = \Sigma c_i y_i / \Sigma c_i$ with $c_i = n_i / [1 + (n_i - 1)\cdot \rho]$ has smaller variance than the sample mean $\tilde{\mu}_p = \Sigma n_i y_i / \Sigma n_i$.

Our approach to this problem combines three statistical methodologies: Empirical Bayes (Morris 1983), generalized linear models (GLIM) and quasi-likelihoods (Nelder and McCullagh 1983), and parametric bootstrap (Efron 1982a and b, Hill 1985). We are primarily concerned with four issues. First, in Section 2 we discuss an Empirical Bayes model for unbalanced dichotomous survey data having positive intra-interviewer correlation. In Section 5, we apply the model to data collected in an experimental random digit dial (RDD) telephone survey. Our approach was able to handle some of the special situations which arise in telephone surveys, but which usually do not arise in personal visit surveys. For example, interviewer sample sizes are often widely discrepant in telephone surveys, since some interviewers work more productive shifts than others; by contrast, interviewer assignments are more easily planned to have nearly equal size in personal visit surveys.

Second, in Section 3 we discuss the estimation of the intra-interviewer correlation coefficient using a standard ANOVA estimator and a maximum quasi-likelihood (MQL) estimator. For two RDD survey items, we compared these estimators using a parametric bootstrap, and were surprised to find the ANOVA estimator outperforming the MQL estimator.

Third, also in Section 3 we discuss the estimation of the population mean μ_p. As suggested already, an estimate of ρ will allow us to improve upon the usual estimator of μ_p.

Fourth, in Section 4 we discuss a technique for identifying discrepant interviewers; i.e., those who make a large contribution to the interviewer variance. Closer supervision of interviewers should be possible in a centralized telephone interviewing facility than in usual field surveys. In fact, "better interviewer control" is often given among the arguments in favor of telephone interviewing. Therefore, our technique for identifying discrepant interviewers should be of special interest to managers of a telephone survey. The Empirical Bayes approach we have taken leads naturally to an interviewer ranking, allowing identification of discrepant interviewers.

2. AN EMPIRICAL BAYES MODEL FOR UNBALANCED DICHOTOMOUS DATA WITH POSITIVE INTRA-INTERVIEWER CORRELATION

The random variables y_{ij} from Section 1 will now be defined specifically for the dichotomous data model. Let $y_{ij} = 1$ if the j^{th} unit of interviewer i's assignment belongs to some category, and $y_{ij} = 0$ otherwise, $i = 1,..., k$; $j = 1, ..., n_i$. Let $p_i = E[y_{ij}|\, i]$ be the probability that interviewer i classifies a randomly chosen unit as being in the category. Given i, we assume that y_{ij} and $y_{ij'}$ are independent; hence $y_{ij}|p_i \sim Bernoulli(p_i)$. If we let $y_i = \sum_j y_{ij} / n_i$ denote the observed proportion for interviewer i, then

$$y_i \mid p_i \overset{ind}{\sim} Bin(n_i, p_i)/ n_i, \qquad (2.1)$$

so that $E(y_i \mid p_i) = p_i$, $Var(y_i \mid p_i) = p_i(1 - p_i) / n_i$.

Now we further assume that the p_i's are independent random variables, as they would be if the interviewers participating in the survey were a sample from an infinite population of interviewers. Specifically, we assume the p_i's are a random sample from a beta distribution having mean μ_p and variance $\rho\mu_p(1-\mu_p)$, which we denote as

$$p_i \sim \text{Beta}[\ \mu_p, \rho\mu_p(1-\mu_p)], \qquad (2.2)$$

$0 \leq \mu_p \leq 1, 0 \leq \rho \leq 1$.

Marginally, the y_i's have beta-binomial distributions with $\text{Var}(y_i) = \rho_i\mu_p(1-\mu_p)/n_i$ where $\rho_i = 1 + (n_i - 1)\rho$, so we write

$$y_i \sim \text{BB}[\ \mu_p, \rho_i\mu_p(1-\mu_p)/n_i]. \qquad (2.3)$$

In the generalized linear model literature (Nelder and McCulloch, 1983), ρ is called a dispersion parameter, but since $\text{Cov}(y_{ij}, y_{ij'}) = \rho\mu_p(1-\mu_p)$ and $\text{Var}(y_{ij}) = \mu_p(1-\mu_p)$ imply $\text{Corr}(y_{ij}, y_{ij'}) = \rho$, ρ is also the intra-interviewer correlation defined in Section 1. This model for extra-binomial variation was one of two explored by Williams (1982).

3. ESTIMATING THE INTERVIEWER CORRELATION AND THE POPULATION MEAN

In this section, we describe two ways to estimate ρ and μ_p. First, we describe the maximum quasi-likelihood estimators, then we describe the usual ANOVA estimators. The mean-variance relationship given in (2.3) results in the following extended quasi-likelihood:

$$q(\mu_p, \rho) = -(1/2)\Sigma\, d_i(\mu_p)/\rho_i - (1/2)\Sigma \log(\rho_i),$$

where $d_i(\mu_p)$ is the i^{th} deviance component,

$$d_i(\mu_p) = \begin{cases} -2n_i \log(1-\mu_p), \text{ if } y_i = 0 \\ 2n_i(y_i \log(y_i/\mu_p) + (1-y_i)\cdot \\ \quad \log[(1-y_i)/(1-\mu_p)]), \quad 0 < y_i < 1 \quad (2.3) \\ -2n_i\log(\mu_p), \text{ if } y_i = 1. \end{cases}$$

To find maximum quasi-likelihood (MQL) estimates of (μ_p, ρ) we set the partial derivatives of q to zero and solve. We obtain $\mu_p = \Sigma c_i y_i / \Sigma c_i$, $c_i = n_i/\hat{\rho}_i$ and $\hat{\rho} = \Sigma w_i S_i/\Sigma w_i$, $w_i = (n_i-1)^2/\hat{\rho}_i^2$ and $S_i = (d_i(\hat{\mu}_p)-1)/(n_i-1)$, which must be calculated iteratively. The following simple algorithm can be used to calculate $(\hat{\mu}_p, \hat{\rho})$:

(0) Fix a starting value for $\hat{\rho}$.
Then calculate (1) $\hat{\rho}_i = 1 + (n_i - 1)\hat{\rho}$
(2) $c_i = n_i / \hat{\rho}_i$
(3) $w_i = (n_i - 1)^2 / \hat{\rho}_i^2$,
(4) $\hat{\mu}_p = \Sigma c_i y_i / \Sigma c_i$,
(5) $d_i = d_i(\hat{\mu}_p)$, according to (2.3)
(6) $S_i = (d_i - 1)/(n_i - 1)$
(7) $\hat{\rho} = \Sigma w_i S_i / \Sigma w_i$
(8) Repeat steps (1) - (7) until $(\hat{\mu}_p, \hat{\rho})$ converges.

Asymptotically, as the number of interviewers $k \rightarrow \infty$,

$$\begin{pmatrix} \hat{\mu}_p \\ \hat{\rho} \end{pmatrix} \sim \left[\begin{pmatrix} \mu_p \\ \rho \end{pmatrix}, \begin{pmatrix} \mu_p(1-\mu_p)/\Sigma c_i & 0 \\ 0 & 2/\Sigma w_i \end{pmatrix} \right] (2.4)$$

In Section 1, we saw that the usual estimator of $E(y_{ij}) = \mu_p$ is the sample mean, $\tilde{\mu}_p$. The MQL estimator $\hat{\mu}_p$ is an alternative to $\tilde{\mu}_p$ and has smaller variance, since $\hat{\mu}_p$ is approximately the best linear unbiased estimator of μ_p, as observed in Section 1.

The usual unbiassed estimator of $\sigma_p^2 = \text{Var}(p_i)$, which is obtained by equating sums of squares to their expected values, is

$$\tilde{\sigma}_p^2 = (V_b - V_w) / (\bar{n} - s_n^2/k\bar{n}),$$

where $\bar{n} = \Sigma n_i/k$, $s_n^2 = \Sigma(n_i - \bar{n})^2/(k-1)$, $V_b = \Sigma n_i(y_i - \tilde{\mu}_p)^2/(k-1)$, and $V_w = \Sigma n_i y_i(1-y_i)/k(\bar{n}-1)$ (Kish 1962, p.110). Then ρ may be estimated by

$$\tilde{\rho} = \tilde{\sigma}_p^2/\tilde{\mu}_p(1-\tilde{\mu}_p).$$

In Section 5, we apply these two sets of estimators to two items on the RDD questionnaire. We also give results of a bootstrap comparison of the estimators.

4. IDENTIFYING DISCREPANT INTERVIEWERS

Managers of a survey need objective measures of interviewer performance for purposes of quality control. Response rates and production measures are routinely tabulated for this purpose in many surveys. In this section, we propose a procedure, based on Empirical Bayes methods, which identifies interviewers whose p_i values are discrepant. We cannot say that the recorded responses of these extreme interviewers are less accurate than those of the others, since we have no measure of bias for any observation. However, we do know that these interviewers contribute to the magnitude of σ_p^2 and thus to the total variance of $\tilde{\mu}_p$. The survey manager may be able, for example by monitoring the interviewer, to determine the cause of the discrepancy.

The distributional assumptions given in (2.1) and (2.2) imply by Bayes Theorem that

$$p_i \mid \underset{\sim}{y} \overset{ind}{\sim} \text{Beta}\left[\mu_i^* = (1 - B_i)y_i + B_i\mu_p, \ \frac{\rho n_i}{1+\rho n_i} \ \frac{\mu_i^*(1-\mu_i^*)}{n_i}\right],$$

where $\underset{\sim}{y} = (y_1, ..., y_k)$ and $B_i = (1 - \rho)/[1 + (n_i - 1)\rho]$. The empirical Bayes estimate of p_i is then $\hat{p}_i = (1 - \hat{B}_i)\, y_i + \hat{B}_i\, \hat{\mu}_p$, where $\hat{B}_i = (1 - \hat{\rho}) /[1 + (n_i - 1)\, \hat{\rho}]$, with $\hat{\rho}$ and $\hat{\mu}_p$ determined from (2.3).

Note that $\hat{p}_i$ is shrunk toward $\hat{\mu}_p$ and away from the usual binomial estimator of p_i, y_i. The shrinkage factors, $\hat{B}_i$, differ among interviewers, with the shrinkage being greatest for interviewers having small n_i and smallest for those having large n_i. In a sense, then, the Empirical Bayes estimates, $\hat{p}_i$, correct for the fact that spuriously large or small values of y_i are likely when the number of cases handled by the i^{th} interviewer is small. For that reason, the ordering of interviewers based on their y_i values may not be retained for the $\hat{p}_i$ values, and thus the interviewers identified as extreme by the two methods may differ. These ideas are illustrated by the two examples in Section 5.

5. EXAMPLES

The two examples in this section are from data collected in an experimental RDD telephone survey conducted between April and September of 1982 by the U.S. Bureau of the Census. About 16 interviewers participated in the experiment and their assignments were interpenetrated within shifts so that estimates of the interviewer correlation could be made.

In the first example, p_i is the item response rate for a salary question on the RDD questionnaire. Columns 2 and 3 of Table 1 show the observed proportion of responses for that item (y_i) in each interviewer's assignment, and the assignment size on which that statistic is based (n_i). We call this first example SAL2.

The usual estimates $\tilde{\rho}$ and $\tilde{\mu}_p$ are $\tilde{\rho} = .0441$ and $\tilde{\mu}_p = .7977$. The estimates $\hat{\rho}$ and $\hat{\mu}_p$ are obtained as described in Section 2. They are $\hat{\rho} = .0378$ and $\hat{\mu}_p = .7889$. Columns 4 and 5 of Table 1 give the shrinkage factors, $\hat{B}_i$, and the empirical Bayes estimates, $\hat{p}_i$, of the interviewer response rates. Figure 1 graphically illustrates the shrinkage pattern of the Empirical Bayes estimates. The crossover in the ordering of interviewers is clearly shown. Interviewer 9 is identified on the basis of the Empirical Bayes estimation procedure as having the best item response rate, while interviewer 4, who has the largest y_i value, but the smallest n_i, drops to sixth.

A graphical display like Figure 1 lets the survey manager see immediately which interviewers are giving evidence of discrepant behavior and which are not. Without such a tool, a large number of pairwise tests would be required to identify the extreme interviewers. That procedure is time-consuming and probably would require a statistician's judgment.

The approximate variances of $\hat{\rho}$ and $\hat{\mu}_p$ can be estimated from (2.4) to be

$$\mathrm{var}(\hat{\mu}_p) = \hat{\mu}_p(1 - \hat{\mu}_p) \,/\, \Sigma c_i = .0005547 = (.0236)^2$$

$$\mathrm{var}(\hat{\rho}) = 2 \,/\, \Sigma w_i = .0003494 = (.0187)^2.$$

We can then obtain from (1.1) and (2.4) an estimate of the approximate relative efficiency of the two estimators of μ_p.

$$\mathrm{RE}(\hat{\mu}_p, \tilde{\mu}_p) = (\Sigma c_i)(1 + (\widehat{n-1})\, \hat{\rho}) \,/\, N$$

$$= \mathrm{Var}(\tilde{\mu}_p)/\mathrm{Var}(\hat{\mu}_p) = 1.232.$$

A similar comparison of theoretical relative efficiencies of the two estimators of ρ is not possible since $\mathrm{Var}(\tilde{\rho})$ has not been determined for the unbalanced binomial case.

In order to compare the exact behavior of the two estimates of ρ, we used a parametric bootstrap. The observed values of $\hat{\mu}$ and ρ were used in the prior to simulate p_i^*'s according to $p_i^* \sim \mathrm{Beta}[.7889, (.0378)(.7889)(.2111)]$. Then the y_i^*'s were simulated according to $y_i^*|p_i^* \sim \mathrm{Bin}(n_i, p_i^*)/n_i$ with the n_i's as given in Column 3 of Table 1. The y_i^*'s were used to calculate $\hat{\mu}^*$ and $\hat{\rho}^*$ and $\tilde{\mu}^*$ and $\tilde{\rho}^*$. This procedure was repeated 1000 times and means and standard deviations of the estimators were calculated. The results for three different repetitions of this bootstrap, together with the theoretical values when available, are given in Table 2 for SAL2.

In the second example, p_i is the proportion of households in interviewer i's assignment recorded as having at least one member answering "Keeping house" to the employment status question. For this example $\tilde{\rho} = .0253$ and $\tilde{\mu}_p = .4324$, while $\hat{\rho} = .0262$ and $\hat{\mu}_p = .4439$. Table 3 shows y_i, n_i, $\hat{B}_i$, and $\hat{p}_i$ for the 12 interviewers having responses to this question and Figure 2 gives the shrinkage pattern. We call this example KH3. A similar bootstrap to that described above was run for KH3 and the results are given in Table 4.

Three results seem apparent from this simulation: $\hat{\mu}_p$ is better than $\tilde{\mu}_p$ as an estimator of

Table 1. Data for SAL2

Interviewer i	y_i	n_i	$\hat{B}_i$	$\hat{p}_i$
1	.861	43	.372	.834
2	.732	41	.383	.754
3	.679	28	.476	.731
4	.929	14	.645	.838
5	.884	249	.093	.875
6	.872	39	.394	.839
7	.692	26	.495	.740
8	.695	95	.211	.715
9	.922	116	.180	.898
10	.882	93	.215	.862
11	.653	124	.170	.676
12	.722	79	.244	.738
13	.775	129	.165	.777
14	.869	122	.173	.855
15	.729	155	.141	.737
16	.738	61	.294	.753

Figure 1 Shrinkage Pattern for SAL2

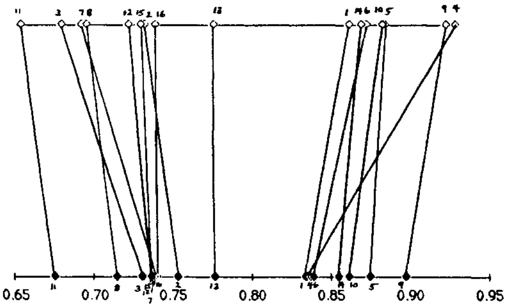

Table 2. Bootstrap for SAL2 based on 1000 repetitions

a. $\hat{\mu}_{obs}$ = .7889

Bootstrap	E($\hat{\mu}$)	E($\tilde{\mu}$)	SD($\hat{\mu}$)	SD($\tilde{\mu}$)	RE-Var($\tilde{\mu}$)/Var($\hat{\mu}$)
1	.7886	.7889	.0231	.0258	1.247
2	.7890	.7894	.0238	.0265	1.240
3	.7892	.7895	.0239	.0260	1.183
Theory	.7889	.7889	.0236	.0262	1.232

b. $\hat{\rho}_{obs}$ = .0378

Bootstrap	E($\hat{\rho}$)	E($\tilde{\rho}$)	SD($\hat{\rho}$)	SD($\tilde{\rho}$)	RE-Var($\tilde{\rho}$)/Var($\hat{\rho}$)
1	.0391	.0376	.0192	.0191	.990
2	.0386	.0372	.0194	.0196	1.021
3	.0394	.0382	.0201	.0203	1.020
Theory	.0378	.0378	.0187	-	-

Table 3. Data for KH3

Interviewer i	y_i	n_i	$\hat{B}_i$	$\hat{p}_i$
1	.609	23	.618	.507
2	.600	25	.598	.507
3	.417	24	.607	.433
4	.500	10	.789	.456
5	.381	134	.217	.394
6	.583	12	.756	.478
7	.438	16	.699	.442
8	.340	53	.412	.383
9	.489	90	.292	.476
10	.476	21	.639	.456
11	.167	30	.553	.320
12	.500	50	.426	.476

Figure 2 Shrinkage Pattern for KH3

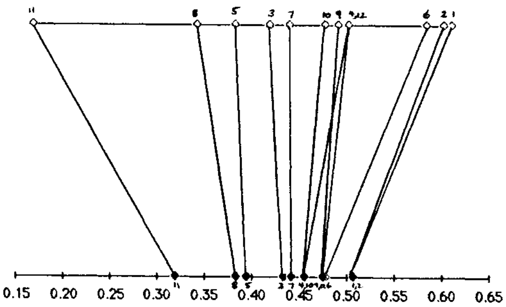

Table 4. Bootstrap for KH3 based on 1000 repetitions

a. $\hat{\mu}_{obs}$ = .4439

Bootstrap	E($\hat{\mu}$)	E($\tilde{\mu}$)	SD($\hat{\mu}$)	SD($\tilde{\mu}$)	RE=Var($\tilde{\mu}$)/Var($\hat{\mu}$)
1	.4446	.4443	.0346	.0374	1.168
2	.4442	.4448	.0353	.0382	1.171
3	.4424	.4434	.0352	.0385	1.196
Theory	.4439	.4439	.0346	.0379	1.200

b. $\hat{\rho}_{obs}$ = .0262

Bootstrap	E($\hat{\rho}$)	E($\tilde{\rho}$)	SD($\hat{\rho}$)	SD($\tilde{\rho}$)	RE-Var($\tilde{\rho}$)/Var($\hat{\rho}$)
1	.0295	.0272	.0260	.0231	.789
2	.0290	.0276	.0265	.0233	.773
3	.0272	.0258	.0260	.0234	.810
Theory	.0262	.0262	.0228	-	-

μ_p, $\tilde{\rho}$ is better than $\hat{\rho}$ as an estimate of ρ, and the bootstrap variance of $\hat{\rho}$ is larger than the asymptotic variance given in Section 3. The first result follows our intuition, but the second and third results were both surprises. We offer several possible explanations for these results: (1) the asymptotic variance for $\hat{\rho}$ should be increased to account for substantial statistical curvature; (2) (related to (1)) since the quasi-likelihood for ρ is that of a curved exponential family, a more appropriate measure of variability is the conditional variance, with conditioning on the appropriate ancillary statistic; (3) $\hat{\rho}$ may improve with larger values of k; (4) the KH3 example may be atypical; (5) estimates of ρ based on deviance may not be as well-behaved as estimates based on sums of squares. Whatever the explanation, for now, we recommend using $\tilde{\rho}$ to estimate ρ and $\hat{\mu}_p$, with $\tilde{\rho}$ replacing $\hat{\rho}$, to estimate μ_p.

Further investigation is needed to determine a good method for accomodating in the model fixed effects, in addition to the random effect caused by the interviewer. For example, respondents reached at different times of the day tend to differ in their responses to some items. Therefore, interviewers working different shifts may differ with respect to their mean responses, but the source of the variability is due to a fixed effect (shift) rather than to the interviewer himself. Other effects which might affect responses are interviewer experience or case priority level. These kinds of differences among interviewers or case types should not be reflected in the interviewer variance. Williams (1982) suggests two models, each of which allows fixed effects in addition to the random effect causing the extra-binomial variation.

REFERENCES

1. Efron, B. (1982a) "The Jackknife, the Bootstrap, and Other Resampling Plans," SIAM - CBMS, 38.
2. Efron, B. (1982b) "Maximum Likelihood and Dȩcision Theory," Annals of Statistics, 7, 1 - 26.
3. Hill, J.R. (1985) "Statistics: An Empirical Bayes Approach," Technical Report #21, Center for Statistical Sciences, The University of Texas at Austin.
4. Kish, Leslie (1962) "Studies of Interviewer Variance for Attitudinal Variables," Journal of the American Statistical Association, 57, 92 - 115.
5. McCullagh, P. and Nelder, J.A. (1983) Generalized Linear Models, Chapman and Hall, NY.
6. Morris, C.N. (1983) "Parametric Empirical Bayes Inference: Theory and Applications (with discussion)", Journal of the American Statistical Association, 78, 47 - 65.
7. Williams, D.A. (1982) "Extra-binomial variation in Logistic Linear Models," Applied Statistics, 31, 144 - 148.