AN APPLICATION OF CATEGORICAL DATA ANALYSIS TO THE NATIONAL HEALTH INTERVIEW SURVEY

Dwight B. Brock, National Center for Health Statistics
Daniel H. Freeman, Jr.,* Jean L. Freeman,* and Gary G. Koch
University of North Carolina

1. Introduction

One approach to the analysis of data from complex sample surveys is the weighted least squares methodology developed by Grizzle, Starmer and Koch (GSK) (1969) to deal with categorical data. The procedure was modified by Koch, Freeman and Freeman (KFF) (1975) to account for some of the effects of sample design on the statistics being analyzed (such as the sampling variance). However, it was not possible at the time to investigate all the factors which might influence the outcome of the analysis. To further study the feasibility of using this technique to analyze survey data, an empirical investigation has recently been conducted using two sets of physician visit data from the Health Interview Survey (HIS) of the National Center for Health Statistics. This paper presents some of the results of this study.

The objective of the study was to test the effects of the following factors on the outcome of the analyses using the GSK methodology:

1) two methods of estimating covariance matrices of ratio estimates using balanced repeated replication (BRR),

2) the influence of poststratification in the ratio estimation procedure on the standard errors of ratio statistics,

3) the assumption of zero covariances among the statistics.

Once these questions were examined, models were fitted using the GSK methodology, and tables of predicted values of the parameters of interest were produced, along with the estimated standard errors of these predicted values. Inferences concerning the parameters of the fitted models were made using Wald (1943) asymptotic χ^2 statistics, which are based on large-sample multivariate normality of the ratio statistics estimated from the complex sample.

2. The Experiment

The balanced repeated replication (BRR) method of variance estimation as described by McCarthy (1966) and Kish and Frankel (1970), can be used in several ways to estimate the covariance matrix of sample statistics. In this experiment Taylor series (TS) approximations (as described in Forthover and Koch (1973)) of the variances and covariances of the ratios were computed using BRR estimates of the variances of the numerator and denominator of each ratio (and the covariance between numerator and denominator). These Taylor series approximations were then compared to estimates of variances and covariances obtained by direct application of BRR to the ratios themselves. The latter method is called the replicated ratio (RR) method.

The second set of comparisons examines variances and inferences for poststratified ratio estimates versus nonpoststratified ratio estimates, for both the Taylor series method and the replicated ratio method of estimating the covariance matrix.

The last set of comparisons looks at the covariances among domains in the crossclassifications. First these covariances are estimated using each of the above four methods and the resulting inferences are compared to the inferences resulting from the assumption of zero covariances among domains.

The data used in the experiment were taken from the 1973 Health Interview Survey. This survey follows a complex multistage stratified probability sample design with ratio adjustment for nonresponse and poststratification to Census Bureau estimates of levels of population in 60 age-sex-race classes. The response variable for the first part of the experiment was the ratio (R) of physician visits (PV) to population (P) in each of 16 cells of an age-sex-race crossclassification. The age ranges were 0-16, 17-44, 45-64, and 65 and over. Sex categories were male and female, and race categories were white and other. For the second part of the experiment the response was again PV/P, this time crossclassified by family income ($0-4,999, 5,000-14,999, 15,000 and over), education of head of household (less than high school; high school; more than high school), and residence (SMSA; non-SMSA).

3. The Age, Sex, Race Crossclassification

The observed estimates for this data set are shown in Table 4. In the poststratified case, since the poststratification is done on the same points as the crossclassification, the denominators of the ratios are constant, and the Taylor series estimates are identical to the replicated ratios. The comparison between TS and RR for nonpoststratified data is shown in Table 1, and little difference is observed.

The effect of poststratification for both TS and RR were examined using the Wald (1943) statistics for total variation and the mean standard errors over the 16 cells in the age-sex-race table. In each case the values for the poststratified data were slightly smaller than for the nonpoststratified, the largest difference being less than four percent.

---

*Present address: Yale University

To examine the zero covariance assumption we first computed the correlation matrix which is shown in Table 2. This gives us some evidence that our covariances are neither all zero nor all positive. Additional comparisons were made using fitted models with covariances estimated and with covariances assumed to be zero (Table 3). We concluded that though the zero covariance assumption was not serious in this case, the substantial reduction in the Q statistics (36% in this case) could cause misleading inferences to be made with other data sets.

Finally, models were fitted for the purpose of producing predicted values and their fitted standard errors. These values are shown in Table 4 along with the original observations and their standard errors.

4. The Income, Education, Residence Cross-classification

The experiment described above was repeated for the observed estimates in Table 9 because it was felt that the conclusions based on the age, sex, race classification may have been limited in two ways. First, age, sex, and race are not believed to be subject to serious response error. Secondly, since poststratification in the HIS is done on precisely the same (age, sex, and race) variables, the effect of poststratification may have been "washed out." In fact the poststratified estimates in the age, sex, race example turned out to be linear sample statistics, which are known to have well-behaved BRR estimates of variance (McCarthy (1966)).

The experimental comparisons from the income, education, and residence data set are given in Tables 5 through 9. The empirical evidence shown here confirms the results obtained from the previous data set. That is, the Taylor series and replicated ratio methods give similar estimates of covariance matrices of ratio statistics. Secondly, the poststratification adjustment does not appear to have an effect on the variances or inferences concerning these ratio statistics. Finally, the assumption of zero covariance among classification domains could have an effect on the inference procedure, since it deflates the total variation as shown by the Wald statistics. Again fitted values and standard errors are given in Table 9 along with the original observations.

TABLE 1

COMPARISON OF ESTIMATED STANDARD ERRORS FOR RR AND TS PROCEDURES. NONPOSTSTRATIFIED AGE x SEX x RACE CLASSIFICATION OF PV/P. COVARIANCES ESTIMATED.

Age	Estimate	Male		Female	
		White	Other	White	Other
<17	RR	0.1244	0.2113	0.1121	0.2215
	TS	0.1242	0.2107	0.1120	0.2210
	RR÷TS	1.0012	1.0027	1.0009	1.0021
17-44	RR	0.0978	0.2340	0.1098	0.3277
	TS	0.0978	0.2333	0.1099	0.3290
	RR÷TS	1.0003	1.0029	0.9996	0.9961
45-64	RR	0.1637	0.4516	0.1530	0.4286
	TS	0.1642	0.4447	0.1530	0.4288
	RR÷TS	0.9973	1.0157	1.0004	0.9994
65+	RR	0.2318	0.9602	0.1907	0.6458
	TS	0.2308	0.9538	0.1906	0.6388
	RR÷TS	1.0043	1.0067	1.0005	1.0109

TABLE 2

CORRELATION MATRIX FOR REPLICATED RATIOS
AGE x SEX x RACE
POSTSTRATIFIED

		<17				17-44				45-64				65+			
		MW	MO	FW	FO	MW	MO	FW	FO	MW	MO	FW	FO	MW	MO	FW	FO
<17	MW	1.0000															
	MO	0.1641	1.0000														
	FW	0.0479	0.0008	1.0000													
	FO	0.1432	0.4356	0.0069	1.0000												
17-44	MW	0.0366	0.0381	0.2661	0.0436	1.0000											
	MO	0.0297	-0.0788	-0.0261	-0.2764	-0.2207	1.0000										
	FW	0.0783	0.0212	0.0905	0.3000	0.1333	-0.3191	1.0000									
	FO	0.3057	0.1771	0.0205	-0.1001	-0.2565	0.3120	-0.2617	1.0000								
45-64	MW	-0.0011	0.0213	-0.1728	-0.1091	-0.0789	0.0662	-0.0453	0.2913	1.0000							
	MO	0.0207	0.0826	-0.0844	0.0185	0.0406	-0.3070	-0.0902	0.1532	0.0703	1.0000						
	FW	0.1588	0.0644	0.2131	0.0168	0.0338	0.0657	-0.0082	0.0262	0.0628	-0.0050	1.0000					
	FO	0.1473	0.0429	-0.0100	-0.1767	-0.1481	0.2177	-0.1040	0.1794	0.0667	0.0711	0.0295	1.0000				
65+	MW	-0.0290	0.0231	-0.1422	-0.0955	-0.0516	-0.1971	-0.0697	0.0965	0.2257	0.1075	0.0901	-0.1481	1.0000			
	MO	-0.0673	0.2199	-0.0602	0.1418	0.0098	-0.1332	-0.0129	-0.2517	0.0351	-0.0081	-0.1437	0.2770	-0.1505	1.0000		
	FW	0.0341	0.0734	0.0542	-0.0316	-0.1282	0.0325	-0.0142	0.4524	0.1089	-0.0521	-0.2073	0.0663	0.1632	-0.0575	1.0000	
	FO	-0.0720	0.0863	-0.1646	0.1493	-0.0731	-0.1455	0.0192	-0.0064	0.0649	0.1994	0.0458	0.1314	0.0679	0.2452	-0.1227	1.0000

TABLE 3

EFFECTS OF ZERO COVARIANCE ASSUMPTIONS ON FINAL MODELS
NONPOSTSTRATIFIED REPLICATED RATIOS
PHYSICIAN VISITS/POPULATION CLASSIFIED BY AGE, SEX, AND RACE

Covariances Estimated	Interdomain Covariances Zero
Parameter Vector $\underset{\sim}{b}$ and Estimated Covariance Matrix $\underset{\sim}{V}(\underset{\sim}{b})$	
$\underset{\sim}{b} = \begin{bmatrix} 6.3791 \\ -0.8316 \\ 0.4967 \end{bmatrix}$; $\underset{\sim}{V}(\underset{\sim}{b}) = \begin{bmatrix} 4.5085 & -1.3265 & 2.0230 \\ -1.3265 & 0.6487 & -0.8717 \\ 2.0230 & -0.8717 & 11.371 \end{bmatrix} \times 10^{-3}$	$\underset{\sim}{b} = \begin{bmatrix} 6.4167 \\ -0.8436 \\ 0.5304 \end{bmatrix}$; $\underset{\sim}{V}(\underset{\sim}{b}) = \begin{bmatrix} 5.0859 & -1.6784 & 1.8516 \\ -1.6784 & 0.8806 & -0.6110 \\ 1.8516 & -0.6110 & 11.802 \end{bmatrix} \times 10^{-3}$
Tables of Variation	

Source	d.f.	Q	% Total
Model	2	1103.90	98.65
Error	13	15.07	1.35
Total	15	1118.97	100.00

Source	d.f.	Q	% Total
Model	2	808.31	98.51
Error	13	12.24	1.49
Total	15	820.55	100.00

Ratios: (Covariance Estimated) ÷ (Covariance = 0)

Parameters: $\begin{bmatrix} 0.994 \\ 0.986 \\ 0.936 \end{bmatrix}$, Covariance Matrix: $\begin{bmatrix} 0.886 & 0.790 & 1.093 \\ 0.790 & 0.737 & 1.427 \\ 1.093 & 1.427 & 0.963 \end{bmatrix}$

Tables of Variation: Model 1.366
Error 1.231
Total 1.364

TABLE 4

OBSERVED AND FITTED VALUES AND MODEL VARIATION FOR PV/P CLASSIFIED BY AGE, SEX, AND RACE. POSTSTRATIFIED DATA. COVARIANCES ESTIMATED.

Table of Values

Age		Sex Race Class: Male White	Male Other	Female White	Female Other
<17	Observed	4.6549	3.0635	4.1199	3.0648
	Fitted	4.7182	3.0622	3.8902	3.0622
	(Obs. S.E.)	(0.1257)	(0.2092)	(0.1108)	(0.2229)
	Fitted S.E.	0.0426	0.0664	0.0495	0.0664
17-44	Observed	3.7129	3.0170	6.3649	6.3755
	Fitted	3.8902	3.0622	6.3743	6.3743
	(Obs. S.E.)	(0.0978)	(0.2298)	(0.1135)	(0.3203)
	Fitted S.E.	0.0495	0.0664	0.0671	0.0671
45-64	Observed	4.8232	4.6611	5.9242	7.0655
	Fitted	4.7182	4.7182	5.8818	6.3743
	(Obs. S.E.)	(0.1629)	(0.4517)	(0.1554)	(0.4242)
	Fitted S.E.	0.0426	0.0426	0.1105	0.0671
65+	Observed	5.8624	8.0478	6.9463	6.2122
	Fitted	5.8818	6.8868	6.8868	6.3743
	(Obs. S.E.)	(0.2288)	(0.9455)	(0.1876)	(0.6514)
	Fitted S.E.	0.1105	0.1399	0.1399	0.0671

Analysis of Variation Table

Source	d.f.	Q	Contrast	S.E.	% Total Q
Model	2	1059.76			98.52
b_1	1	1036.29	-0.8280	0.0257	
b_2	1	21.32	0.4925	0.1067	
$b_1 + 2b_2 = 0$	1	0.57	0.1570	0.2074	
Error	13	15.93			1.48
Total	15	1075.69			

TABLE 5

COMPARISON OF ESTIMATED STANDARD ERRORS
FOR REPLICATED RATIOS (RR) AND TAYLOR SERIES APPROXIMATION (TS),
INCOME BY RESIDENCE BY EDUCATION CLASSIFICATION
OF PHYSICIAN VISITS/POPULATION
(COVARIANCES ESTIMATED)

Education	Estimate	Income and Residence Class: 0-4,999 SMSA	0-4,999 Non-SMSA	5,000-14,999 SMSA	5,000-14,999 Non-SMSA	15,000 and over SMSA	15,000 and over Non-SMSA
Poststratified							
<HS	RR	0.1796	0.2618	0.1286	0.1541	0.2530	0.3748
	TS	0.1798	0.2607	0.1285	0.1535	0.2528	0.3755
	RR÷TS	0.9959	1.0042	1.0008	1.0039	1.0008	0.9981
HS	RR	0.4112	0.4378	0.1746	0.1935	0.1799	0.3279
	TS	0.4098	0.4380	0.1746	0.1934	0.1802	0.3248
	RR÷TS	1.0034	0.9995	1.0000	1.0005	0.9983	1.0095
>HS	RR	0.4925	0.5809	0.1868	0.2921	0.1628	0.3128
	TS	0.4885	0.5622	0.1861	0.2915	0.1630	0.3110
	RR÷TS	1.0082	1.0333	1.0038	1.0021	0.9988	1.0058

TABLE 6

EFFECTS OF POSTSTRATIFICATION (PS) VERSUS NONPOSTSTRATIFICATION (NPS).
INCOME (I) BY RESIDENCE (R) BY EDUCATION (E) CLASSIFICATION.
PHYSICIAN VISITS/POPULATION.
COVARIANCES ESTIMATED.

Effects on Standard Errors: S.E.(PS) ÷ S.E.(NPS)

Education	Income and Residence Class: 0-4,999 SMSA	0-4,999 Non-SMSA	5,000-14,999 SMSA	5,000-14,999 Non-SMSA	15,000 and over SMSA	15,000 and over Non-SMSA
	Replicated Ratios					
<HS	1.0096	0.9966	0.9954	1.0013	0.9953	1.0037
HS	1.0054	1.0051	1.0046	1.0010	1.0000	0.9994
>HS	1.0026	0.9983	1.0086	1.0000	0.9994	1.0010

TABLE 7

CORRELATION MATRIX FOR INCOME BY RESIDENCE BY EDUCATION
REPLICATED RATIOS. POSTSTRATIFIED DATA.

Income			0-4,999						5,000-14,999						15,000 and over					
	Residence		SMSA			Non-SMSA			SMSA			Non-SMSA			SMSA			Non-SMSA		
		Education	<HS	HS	>HS	<HS	HS	>HS	<HS	HS	>HS	<HS	HS	>HS	<HS	HS	>HS	<HS	HS	>HS
0-4,999	SMSA	<HS	1.000																	
		HS	0.030	1.000																
		>HS	0.097	-0.286	1.000															
	Non-SMSA	<HS	0.043	0.072	-0.056	1.000														
		HS	0.090	0.078	-0.057	0.134	1.000													
		>HS	-0.096	0.074	-0.087	0.185	0.200	1.000												
5,000-14,999	SMSA	<HS	0.116	0.250	0.071	0.114	-0.085	-0.108	1.000											
		HS	-0.138	-0.076	0.248	-0.051	-0.240	0.044	0.348	1.000										
		>HS	-0.125	0.094	-0.144	-0.016	0.124	0.043	-0.126	-0.113	1.000									
	Non-SMSA	<HS	0.173	-0.058	-0.200	0.080	0.055	-0.042	-0.040	-0.191	0.073	1.000								
		HS	-0.105	-0.140	-0.058	0.082	0.026	0.064	0.018	-0.080	0.064	0.312	1.000							
		>HS	-0.128	0.063	-0.038	0.002	-0.029	0.120	0.030	0.126	-0.016	0.091	-0.029	1.000						
15,000 and over	SMSA	<HS	-0.067	0.155	-0.063	0.043	-0.138	0.050	0.200	0.250	-0.197	-0.076	-0.023	0.160	1.000					
		HS	-0.128	-0.044	0.010	0.083	0.010	0.032	-0.230	-0.124	0.172	0.023	0.031	0.099	-0.100	1.000				
		>HS	-0.292	0.075	-0.252	0.026	-0.034	-0.033	-0.064	-0.086	0.030	-0.017	-0.086	0.017	-0.132	0.209	1.000			
	Non-SMSA	<HS	-0.123	-0.171	0.006	-0.017	0.088	-0.015	0.034	0.185	0.010	0.024	0.084	0.085	0.092	-0.068	0.004	1.000		
		HS	0.052	0.024	0.168	0.007	0.032	0.165	0.067	0.075	-0.042	-0.051	0.003	-0.136	-0.052	0.047	-0.087	-0.039	1.000	
		>HS	0.030	-0.077	-0.050	0.060	0.038	0.044	0.120	-0.080	-0.019	0.137	0.108	0.185	-0.081	0.021	0.173	-0.066	-0.078	1.000

TABLE 8

EFFECTS OF ZERO COVARIANCE ASSUMPTION ON FINAL MODELS
POSTSTRATIFIED REPLICATED RATIOS
PHYSICIAN VISITS/POPULATION.
CLASSIFIED BY INCOME, RESIDENCE, AND EDUCATION.

Covariance Estimated	Interdomain Covariances Assumed to be Zero
Parameter Vector $\underset{\sim}{b}$ and Estimated Covariance Matrix $\underset{\sim}{V}(\underset{\sim}{b})$	
$\underset{\sim}{b} = \begin{bmatrix} 5.0920 \\ 0.2777 \end{bmatrix}$	$\underset{\sim}{b} = \begin{bmatrix} 5.0748 \\ 0.2760 \end{bmatrix}$
$\underset{\sim}{V}(\underset{\sim}{b}) = \begin{bmatrix} 2.5259 & -0.2070 \\ -0.2070 & 0.3882 \end{bmatrix} \times 10^{-3}$	$\underset{\sim}{V}(\underset{\sim}{b}) = \begin{bmatrix} 2.6685 & 0.0926 \\ 0.0926 & 0.5031 \end{bmatrix} \times 10^{-3}$
Tables of Variation	

Source	d.f.	Q	% Total
Model	1	198.64	91.60
Error	16	18.21	8.40
Total	17	216.85	100.00

Source	d.f.	Q	% Total
Model	1	151.36	91.82
Error	16	13.49	8.18
Total	17	164.85	100.00

Ratios: (Covariance Estimated) ÷ (Covariance = 0)

Parameters: $\begin{bmatrix} 1.003 \\ 1.006 \end{bmatrix}$, Covariance Matrix: $\begin{bmatrix} 0.947 & -2.235 \\ -2.235 & 0.772 \end{bmatrix}$

Tables of Variation: Model 1.312
Error 1.350
Total 1.315

5. Summary and Conclusions

This paper has dealt with the problem of fitting linear models to complex survey data using the weighted least squares approach of Grizzle, Starmer and Koch (1969). The experimental investigation reported here found no differences in the Taylor series and replicated ratio methods of estimating covariance matrices of ratio statistics. Similarly, no differences were found between the poststratified and non-poststratified estimates. This, however, may be due to the large size of the HIS sample (120,000 cases) or to the fact that the effect of post-stratification may have been eliminated by examining the ratio of two estimates, each of which had been poststratified before the ratio was computed. Further research on this point is indicated. The study did show that the assumption of zero covariance among domains in the crossclassification produced inflated estimates of variance and substantially reduced levels of variation in the fitted models. Additional work is needed to explore the effects of these factors on variables with different response characteristics, wider ranges of values, and on data from smaller sample sizes.

TABLE 9

OBSERVED AND FITTED VALUES AND MODEL VARIATION FOR PV/P CLASSIFIED BY INCOME, EDUCATION AND RESIDENCE. POSTSTRATIFIED DATA. COVARIANCES ESTIMATED.

Table of Values

Family Income		Residence-Education Class					
		SMSA			Non-SMSA		
		<HS	HS	>HS	<HS	HS	>HS
0-4,999	Observed	6.1475	6.1736	6.3065	5.0770	5.3602	4.5846
	Fitted	5.9250	5.9250	5.9250	5.3697	5.3697	5.3697
	(Obs. S.E.)	(0.1796)	(0.4112)	(0.4925)	(0.2618)	(0.4378)	(0.5809)
	Fitted S.E.	0.0691	0.0691	0.0691	0.0500	0.0500	0.0500
5,000 - 14,999	Observed	4.7348	4.9812	6.0771	4.1442	4.3202	5.0603
	Fitted	4.8143	4.8143	5.9250	4.2589	4.2589	5.3697
	(Obs. S.E.)	(0.1286)	(0.1746)	(0.1868)	(0.1541)	(0.1935)	(0.2921)
	Fitted S.E.	0.0577	0.0577	0.0691	0.0852	0.0852	0.0500
15,000 and up	Observed	4.8245	4.7031	5.6562	4.4177	4.4929	4.4798
	Fitted	4.8143	4.8143	5.9250	4.2589	4.2589	4.2589
	(Obs. S.E.)	(0.2530)	(0.1799)	(0.1628)	(0.3748)	(0.3279)	(0.3128)
	Fitted S.E.	0.0577	0.0577	0.0691	0.0852	0.0852	0.0852

Analysis of Variation Table

Source	d.f.	Q	Contrast	S.E.	% Total Q
Model	1	198.64	0.2777	0.0197	91.60
Error	16	18.21			8.40
Total	17	216.85			

ACKNOWLEDGMENTS

This research was supported in part by the National Institutes of Health (Grants GM-70004-05 and HD-00371). The authors would like to thank E. Earl Bryant and Daniel G. Horvitz for discussions which were helpful in carrying out this study.

REFERENCES

Forthover, R.N. and Koch, G.G. (1973). An analysis for compounded functions of categorical data, Biometrics 29, 143-157.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models, Biometrics 25, 489-503.

Kish, L. and Frankel, M.R. (1970). Balanced repeated replications for standard errors, J.A.S.A. 65, 1071-1094.

Koch, G.G., Freeman, D.H. and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys, Int'l. Stat. Rev. 43, 55-74.

McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys, Series 2 - No. 14, National Center for Health Statistics.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Amer. Math. Soc. 54, 426-482.