A STRATEGY FOR THE ANALYSIS OF MULTIPLE AREA STUDIES

Daniel H. Freeman, Jr., Robert W. Makuch, and Jan A. J. Stolwijk, Yale University School of Medicine

Many studies are designed in such a way that detailed information on individuals is collected by the use of a personal interview or examination. This process is then repeated in a variety of different geographic areas. The objective of such studies is to assess the variation of personal social and/or health attributes across a variety of environmental or ecological conditions. Thereby appropriate associations among the study variables may be evaluated. Studies of this type may be used to generate hypotheses concerning causal relationships and to support a variety of policy decisions. It should be recognized that such studies cannot actually test causal hypotheses, but replication of the results provides reassurance to the investigators about the phenomena under study. For these reasons it is important that statisticians examine the methodological issues associated with integration of individual measurement data and area wide aggregate data.

An aspect of this problem which is not widely discussed is the appropriate methodology when the individual measurements are categorical and the aggregate measurements are continuous in nature. For example, the individual measurements may be smoking status, sex, area of residence and size of urban area and the area wide measurements may be concentration of pollutants. In Table 1 we have precisely this type of data where three different pollutants are measured in micrograms per stere or cubit meter. The details concerning the collection and alternative analyses of the data are found in other sources (Berman(1976) and U.S. Environmental Protection Agency (1974)).

Table 1. Prevalence of chronic bronchitis per 100 population and sample size by smoking status, gender, area of study, and pollutant exposure.

Smoking Status and Gender		Area of study: Western Metropolitan				Western Non-Metropolitan				Eastern Metropolitan				Eastern Non-Metropolitan		
Never Smoked	Female	2.3	2.0	4.7	5.2	1.4	0.5	1.1	3.6	1.5	2.0	7.5	4.9	2.1	4.0	3.5
	n	755	755	772	667	440	207	94	337	333	197	411	529	384	202	344
	Male	3.0	3.6	2.3	6.8	1.1	0.0	4.9	4.9	2.0	4.6	18.0	14.2	2.8	5.2	5.2
	n	396	367	350	265	273	87	41	102	100	174	384	499	214	97	115
Formerly Smoked	Female	5.3	4.0	7.0	7.1	3.0	4.6	0.0	1.8	3.9	3.8	9.0	4.5	1.4	7.8	0.0
	n	75	101	114	84	131	66	27	112	102	144	233	226	140	64	61
	Male	2.6	3.4	5.4	6.0	0.0	5.3	0.0	4.7	5.0	13.9	18.0	18.7	5.7	8.2	5.6
	n	230	177	241	133	244	113	58	127	101	144	222	198	212	98	90
Currently Smokes	Female	17.1	14.7	15.3	22.2	8.7	14.2	13.7	13.8	11.8	13.9	19.8	16.6	7.1	7.0	9.8
	n	214	286	295	212	218	205	95	376	315	267	535	607	128	281	183
	Male	19.9	18.6	20.1	26.8	12.4	20.9	20.0	18.4	19.0	13.9	21.3	22.1	15.0	15.2	17.6
	n	272	311	354	209	209	187	85	250	260	216	492	526	132	287	159
Pollutant		Average Annual Concentration in micrograms per cubic meter (μ g/s)														
Sulphur Dioxide (SO_2)		10.	18.	32.	92.	10.	26.	67.	177.	374.	30.	174.	247.	13.	14.	4.
Total Suspended Particulates (PA)		88.	84.	50.	70.	50.	45.	115.	65.	102.	41.	84.	108.	38.	63.	48.
Suspended Sulphates		3.7	4.7	8.6	15.0	3.3	4.9	7.3	7.2	11.3	10.0	8.6	14.8	5.8	7.8	6.8

Inevitably, the comparability of the measurement across areas is of major concern and there is always the risk of confounding among the variables of interest. Typically, the individual measurements are thought of as blocking variables and the analysis focuses on some response of interest as if it were a continuous and normally distributed random variable. Moreover, it is assumed that the errors or residuals of such a model have essentially constant variance. A final assumption which is almost always made is that individual measurements are made on members of a simple random of some operationally defined population.

D.R. Cox (1970: pp. 16-18) notes that if the underlying probabilities associated with the response of interest lie between 0.2 and 0.8 then the usual least squares analysis will not in general be misleading. However, for data such as in the example this is clearly inappropriate since the observed prevalences range between 0 and 26.8 per cent. Moreover, for groups such as females who do not now smoke the prevalences are strictly less than 10.0 per cent. Cox notes several other difficulties with ordinary least squares analysis:

1. The method of estimation cannot be fully efficient.
2. The predicted values must be restricted to lie between 0 and 1.
3. It is not reasonable to extrapolate the regression equations outside the range of observation because of the obvious approximation being used in the linear equations.

Cox observed that each of these objections may be dealt with for binary variables through use of the logistic transformation. However, he does not examine the problem of what to do when the response is polytomous and ordinal or when the sample is actually based on a complex probability sample.

When the latter two issues are important a more general methodology and strategy of analysis is required. The strategy of Koch, Freeman, and Freeman (1975) (KFF) provides the appropriate framework for analysis. It is based on an elaboration of the method of Grizzle, Starmer, and Koch (1969) (GSK). It has been employed in a previous analysis of multiple area studies by Makuch and Freeman(1976) using the data of Heneley, Jain and Wells (1976). An aspect of the strategy known as modularization (Freeman, Freeman, and Brock (1977)) is appropriate for the analysis of data sets such as in the example. It is important to note that while the data set at hand is binary and thus lends itself to the logistic transformation, this is <u>not</u> a necessary condition for the analysis. If the original data were available it would be possible to use the original scaling (to 7) or alternatively either a ridit or probit scaling of the responses. Rather than re-iterate previously published material the data will be used to illustrate a strategy for the analysis of multiple area studies.

As noted earlier the data are divided in 6 sub-populations according to smoking status and gender. There are a total of 15 areas where these sub-populations were examined and data were collected on the atmospheric concentration of sulphur dioxide, suspended particulates and suspended sulphates. The areas may be broken into two regions and two urban classes, (West, East) and Metropolitan, Other). Notice this characterization leads to an unbalanced design. The logit of prevalance rates for each of six sub-populations was fit to the linear model shown in Table 2a.

This model is relatively straight-forward and "b" corresponds to a "base-line prevalence of chronic bronchitis" in eastern non-metropolitan populations. "R" is the change in bronchitis rates found in the West while "U" is the metropolitan effect. Notice that these two are treated as additive effects on the logistic scale. A significant interaction would have been equivalent to confounding in the data. Alternatively it would mean that at most two pollutants could be examined. The fourth parameter i the effect due to SO_2. There was no evidence of an interaction between SO_2 and either region or urbanity. The remaining four terms correspond to regional effects of particulates (PA) and sulphates (SU). Again there was no evidence of pollutant by urbanity interaction. Using the weighted least squares algorithm, KFF and GSK, leads to parameter estimates which are

1. Fully efficient for large samples (GSK),
2. Can incorporate either the simple or complex random sample design (KFF),
3. Computationally straight forward,
4. Robust against heteroscedasticity (GSK),
5. Available on any scale involving linearizable functions (KFF).

The resulting test statistics are shown for each sub-population or module in <u>Table 2b.</u> The test of fit of the model is non-significant in each module. The region effect is significant in 5 modules, its interaction in 3. Overall there is a significant pollution effect in four modules. This may be broken into its components; showing SO_2 in only one module, sulphates in four, and particulates in two. Moreover the separate regional sulphate and particulate effects are clearly necessary. One may then interpret these tests or more appropriately indices of significance by considering the corresponding estimates shown in <u>Table 2c.</u>

The effects indicate an increase in bronchitis if the estimate is positive. The region effect is generally small but dramatically reduced bronchitis among Western males who have never smoked. The persons in metropolitan areas have elevated rates. SO_2 has relatively little effect. Where the particulate effect is significant it is negative in the West and positive in the East. Conversely sulphates are positive in the West and negative in the East.

The next step in the analysis is to combine the effects across the modules. This was done following the algorithm of Freeman, Freeman and Brock (1977). It is entirely comparable to backwards elimination in regression analysis. This results in the model shown in Tables 3a to 3c. Based on the fit statistic it is evident that the model is quite acceptable. All of the parameters are nominally significant at the 0.05 level. However, if one adjusts the degrees of freedom to reflect the appropriate variation space (shown under total) only WPA becomes non-significant. Interpretations of the parameters are shown in Table 3b and the corresponding module parameter estimates are shown in Table 3c. Either these or the estimates in Table 3a. may be used to generate the approximate response surfaces.

Briefly the analysis indicates that in the West there is no sex differences among non-smokers. The urban dweller generally has an increased prevalence. There is no evidence in these data of an effect due to SO_2. In the West particulates have a small negative effect among non-smokers but sulphates clearly increase bronchitis for all groups. The eastern picture is basically reversed.

Thus in the East it appears that particulates are associated with increased bronchitis prevalence in all groups, and sulphates have an unexplained negative correlation with bronchitis.

ACKNOWLEDGEMENTS

The authors wish to thank Mrs. Bea Zinn and Miss Joann Raccio for their administrative support.

REFERENCES

Grizzle, S.E., Starmer, C.F., and Koch, G.G.(1969). Analysis of categorical data by linear models. Biometrics 25, 489-504.

Henley, N.S., Jain, S.C., and Wells, H.D. (1976). Relative importance of program input and environmental constraints to family planning programs in Haryana, India. Presented at the 1976 Annual Meeting of the Population Association of America (in Montreal).

Makuch, R.W., and Freeman, D.H. (1976). A multiple logit analysis of a family planning system. Soc. Stat. Sect. Proc. Am. Stat. Assoc. 572-577.

Cox, D.R. (1970). The Analysis of Binary Data. London: Methuen Co. LTD.

Koch, G.G., Freeman, D.H., and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. Int. Stat. Rev. 43, 59-78.

Berman, M.D. (1976). The impact of sulphur dioxide pollution on chronic respiratory disease. Unpublished manuscript.

U. S. Environmental Protection Agency (1974). Health Consequences of Sulphur Oxides, Research Triangle Park, N.C.: National Environmental Research Center.

Freemen, D.H., Freeman, J.L., Brock, D.B. (1977). Modularization for the analysis of interactions in complex sample survey data. Proceedings of the 41st Session of the International Statistical Institute. To appear.

Table 2a. Model used within each gender - smoking module

$$X_w b = \begin{bmatrix} 1 & 1 & 1 & 10. & 88. & 0 & 3.7 & 0 \\ 1 & 1 & 1 & 18. & 84. & 0 & 4.7 & 0 \\ 1 & 1 & 1 & 32. & 50. & 0 & 8.6 & 0 \\ 1 & 1 & 1 & 92. & 70. & 0 & 15.0 & 0 \\ 1 & 1 & 0 & 10. & 50. & 0 & 3.3 & 0 \\ 1 & 1 & 0 & 26. & 45. & 0 & 4.9 & 0 \\ 1 & 1 & 0 & 67. & 115. & 0 & 7.3 & 0 \\ 1 & 1 & 0 & 177. & 65. & 0 & 7.2 & 0 \\ 1 & 1 & 0 & 374. & 102. & 0 & 11.3 & 0 \\ 1 & 0 & 1 & 30. & 0 & 41. & 0 & 10.0 \\ 1 & 0 & 1 & 174. & 0 & 84. & 0 & 8.6 \\ 1 & 0 & 1 & 247. & 0 & 108. & 0 & 14.8 \\ 1 & 0 & 0 & 13. & 0 & 38. & 0 & 5.8 \\ 1 & 0 & 0 & 14. & 0 & 63. & 0 & 7.8 \\ 1 & 0 & 0 & 4. & 0 & 48. & 0 & 6.8 \end{bmatrix} \begin{bmatrix} b \\ \\ R \\ \\ U \\ \\ SO_2 \\ \\ WPA \\ \\ EPA \\ \\ WSU \\ \\ ESU \end{bmatrix}$$

Table 2b. Analysis of variation within modules, Q-statistics

Source	df	Never Smoked		Once Smoked		Now Smokes	
		Female	Male	Female	Male	Female	Male
Model	7	47.00*	131.90*	13.46	80.81*	39.77*	20.56*
Region Total	3	19.52*	38.03*	7.96*	33.45*	7.47	7.85*
Interaction	2	13.47*	16.60*	7.76*	2.34	4.04	5.08
Urban Total	1	2.96	5.27*	2.12	7.64*	19.14*	1.13
Pollution Total	5	25.83*	30.20*	7.98	13.94*	9.04	12.41*
SO_2 Total	1	0.03	0.31	0.19	4.99*	0.80	0.40
Particulate Total	2	12.10*	7.97*	4.83	4.61	2.79	5.15
Interaction	1	10.21*	2.08	4.35*	2.22	2.78	3.99*
Sulphate Total	2	13.70*	16.79*	7.26*	0.06	3.55	6.60*
Interaction	1	12.18*	16.41*	7.01*	0.06	3.53	4.33*
Within Module Error	7	8.24	8.11	6.63	6.78	13.06	6.78
Total Variation	14	55.24*	140.01*	20.08	87.59*	52.84*	27.35*
Percent Explained		85.1	94.1	67.0	92.3	75.3	75.2

*Statistic excedes 95-th percentile of corresponding χ^2 distribution

Table 2c. Within module parameter estimates and estimated standard errors.

Module Label	Estimates and Standard Errors					
	Never Smoked		Once Smoked		Now Smokes	
	Female	Male	Female	Male	Female	Male
b	-3.51*	-3.36*	-3.10*	-2.42*	-2.26*	-2.03*
SE	0.48	0.43	0.70	0.41	0.26	0.25
R: Present if west	0.10	-1.88*	-0.32	-0.18	0.16	0.19
SE	0.65	0.83	0.93	0.71	0.34	0.31
U: Present if metro.	0.41	0.57*	0.47	0.63*	0.55*	0.12
SE	0.24	0.25	0.32	0.23	0.12	0.11
SO_2 (10^{-8} g/s)	0.03	0.11	-0.10	0.39*	0.07	-0.04
SE	0.15	0.20	0.22	0.17	0.08	0.07
Western Particulates	-1.31*	1.14	0.30	-1.89*	-0.19	0.12
(10^{-8} g/s) SE	0.62	0.86	1.08	0.89	0.35	0.31
Eastern Particulates	2.50*	2.93	2.49	-0.47	0.66	1.14*
(10^{-8} g/s) SE	1.20	1.08	1.31	0.83	0.50	0.51
Western Sulphates	6.86*	7.26*	5.26	0.94	1.65	4.26*
(10^{-8} g/s) SE	2.69	3.48	5.34	4.51	2.04	1.87
Eastern Sulphates	-17.66*	-16.35*	-18.88*	-0.59	-5.17	-3.27
(10^{-8} g/s) se	6.41	4.51	7.73	4.98	3.16	3.24

*Ratio of estimate squared to variance exceeds 95-th percentile of χ^2 distribution, df = 1.

Table 3a. Final across module analysis

Analysis of Variation Source/Label	Estimate	Standard Error	Degrees of Freedom Total	Net	Q
Model	———	———	47	8	915.66
b	-3.64	0.13	—	—	——
S	1.21	0.13	16	1	91.60
G	0.69	0.06	8	1	128.09
U	0.50	0.07	6	1	59.38
WPA ($\times 10^{-8}$ g/s)	-0.53	0.19	6	1	7.85
EPA ($\times 10^{-8}$ g/s)	0.94	0.12	6	1	57.80
WSU ($\times 10^{-8}$ g/s)	4.65	0.79	6	1	35.03
ESU ($\times 10^{-8}$ g/s)	-17.55	2.33	6	1	56.52
SESU($\times 10^{-8}$ g/s)	13.26	1.75	6	1	57.56
Error	Final Reduction			8	12.25
	Backwards Elimination			19	23.73
	Initial Model			12	5.19
	Within Modules			42	49.60
	Total Error			81	90.77
Total				89	1006.43

Table 3b. Interpretation of final model parameters

Label	Coefficient(s)	Interpretation of effect on bronchitis
b	1	baseline logit - prevalence of chronic bronchitis for western females who are non-metropolitan and have never smoked.
S	1	person now smokes
	0	otherwise
G	1	Eastern male or smoking western male
	0	otherwise
U	1	Metropolitan person other than smoking males
	0	otherwise
WPA	(Particulates x 10^{-8} g/s)	for western non-smokers
	0	othewise
EPA	(Particulates x 10^{-8} g/s) x 3	for eastern never smoked or female ex-smokers
	x 1	for eastern smoker or male ex-smoker
	x 0	otherwise
WSU	(Sulphates x 10^{-8} g/s)	if western person
	0	otherwise
ESU	(Sulphates x 10^{-8} g/s)	if eastern person except male ex-smokers
	0	otherwise
SESU	(Sulphates x 10^{-8} g/s)	if eastern smoker
	0	otherwise

Table 3c. Fitted within module parameter estimates and estimated standard errors based on final model.

	Estimates and Standard Errors					
Module Label	Never Smoked		Once Smoked		Now Smokes	
	Female	Male	Female	Male	Female	Male
b	-3.64	-2.95	-3.64	-2.95	-2.43	-1.73
SE	0.13	0.12	0.13	0.12	0.08	0.07
R: Present if west	0	-0.69	0	-0.69	0	0
SE	0	0.06	0	0.06	0	0
U: Present if metro.	0.50	0.50	0.50	0.50	0.50	0
SE	0.07	0.07	0.07	0.07	0.07	0
SO_2 (10^{-8} g/s)	0	0	0	0	0	0
SE	0	0	0	0	0	0
Western Particulates	-0.53	-0.53	-0.53	-0.53	0	0
(10^{-8} g/s)SE	0.19	0.19	0.19	0.19	0	0
Eastern Particulates	2.81	2.81	2.81	0.94	0.94	0.94
(10^{-8} g/s)SE	0.37	0.37	0.37	0.12	0.12	0.12
Western Sulphates	4.65	4.65	4.65	4.65	4.65	4.65
(10^{-8} g/s)SE	0.79	0.79	0.79	0.79	0.79	0.79
Eastern Sulphates	-17.55	-17.55	-17.55	0	-4.28	-4.28
(10^{-8} g/s)SE	2.33	2.33	2.33	0	1.09	1.09