〈Research Paper〉

Journal of the Korean
Statistcal Society
Vol.12, No.1, 1983

Comparison of Parameter Estimation Methods in the Analysis of Multivariate Categorical Data with Logit Models**

Hae Hiang Song*

ABSTRACT

In fitting models to data, selection of the most desirable estimation method and determination of the adequacy of fitted model are the central issues. This paper compares the maximum likelihood estimators and the minimum logit chi-square estimators, both being best asymptotically normal, when logit models are fitted to infant mortality data. Chi-square goodness-of-fit test and likelihood ratio one are also compared. The analysis infant mortality data shows that the outlying observations do not necessarily result in the same impact on goodness-of-fit measures.

1. Introduction

Data collected from a regionalized project+ was analyzed to study effects of social, biological and health-care factors on infant mortality. During six weeks in September-October 1978, retrospective informations on 23,514 single births during the past ten years (1968—1978) were obtained. Seo (1982) carried out a preliminary analysis of these 23,514 single births by studying mortality rates and suggested the following five variables for a further study: Maternal age, maternal education, family's socioeconomic status,

* Department of Preventive Medicine & Biostatistics, Catholic Medical College

** This work was supported by the Research Fund of the Ministry of Education, Korean Government, 1982.

+ The project was jointly coordinated by Bureau of Census, Economic Planning Board and Population and Family Planning Program, Yonsei University.

infant's sex and the time interval between present and previous child. These variables were selected by balancing several competing considerations. Compatibilty with other infant mortality studies and the need for adequate number of births within each grouping has limited the choice as above.

Each variable is classified into two groups for a multivariate analysis. Maternal age is classified into two groups of 20-29 years and under 20 or over 29 years. Since births were judged to be at medical-obstetric risk if the maternal age falls outside of a certin age interval, the groups above were suggested by medical researchers (Seo (1982), Bross and Shapiro (1982)). Bross and Shapiro (1982) in fact considered groupings of 19-33 years and under 19 or over 33 years. Categories of birth interval are under 2 years and 2 years or over. For the cases of birth from the first pregnancy, medical-obstetric condition is known to be similar to that of a birth with greater birth interval and hence births from the first pregnancy are joined to the latter category.

In previous studies (Neutra et al. (1978), Quick et al. (1981)) an intervention program of prenatal care was evaluated by assessing its effect on infant mortality and hence relative importance of various factors on mortality was not determined. And, Bross and Shapiro (1982) somewhat unjustifiably made a causal assumption that the relationship among several factors can be decomposed into a direct and an indirect component and hence they studied indirectly the relationships of various factors to mortality through a variable of birthweight.

When several factors are substantially interrelated, a model which will allow the simultaneous estimation of the influence of a number of factors is clearly needed. One possibility is using a multiple regression method in which the effect of any factor is expressible as the increase or decrease in the mortality rate. Feldstein (1966) investigated the relationship between binary variables of maternal age, parity and social class to infant mortality by a method of multiple regression. However, studies using the logit model, as in the present study, have an advantage in that they use the relative risk to measure the relationships. In other words, the effect of any factor is expressed as an odds ratio which is just a percentage increase or decrease in the odds on mortality. In the sections following method of analyzing classificatory data by logit models is described.

2. Logit Models

The logit model is most frequently used in studies in which the assessment of the effects of categorical variables on a dichotomous response variable(e.g., survival or death) is of major concern. Table 1 presents a six-dimensional multi-way table of frequencies including the dichotmous response variable of infant's survival. When two-dimensional margins are examined, as a preliminary analysis, by tabulating survived or dead infants for each binary predictor variable, significant associations are found for all predictor variables (all p-values less than 0.05). Although these results are informative, they do not allow one to study the relationship between mortality and each predictor variable while controlling other variables.

Logit formed for the response variable in a simple model, without interaction terms, can be written as follows:

$$l_{ijklm}=\ln\left(\frac{m_{ijklm1}}{m_{ijklm2}}\right)=\alpha+\beta_{1i}x_{1i}+\beta_{2j}x_{2j}+\beta_{3k}x_{3k}+\beta_{4l}x_{4l}+\beta_{5m}x_{5m}, \quad (1)$$

where $\sum\beta_{1i}=\sum\beta_{2j}=\sum\beta_{3k}=\sum\beta_{4l}=\sum\beta_{5m}=0.$

Here, m_{ijklm1} and m_{ijklm2} are the expected cell frequencies and x_1, x_2, x_3, x_4 and x_5 are indicating respectively explanatory variables of maternal age, maternal education, family's socioeconomic status, infant's sex and birth interval in the order presented. The model states that there are additive effects on the log mortality ratio due to explanatory variables.

In situations where one variable is considered as a response variable and the remainder as predictor variables, product-multinomial typically occurs. For each combination of predictor variables, a sample with fixed size $n_{ijklm}(=m_{ijklm1}+m_{ijklm2})$ is chosen and the crossclassication of each sample according to the response variable is determined by a multinomial distribution.

The estimated probability of death therefore can be derived as follows from(1) :

$$\hat{p}_{ijklm}=1-\hat{q}_{ijklm}=\frac{1}{1+\exp\{-(a+b_{1i}x_{1i}+b_{2j}x_{2j}+b_{3k}x_{3k}+b_{4l}x_{4l}+b_{5m}x_{5m})\}}, \quad (2)$$

where $a, b_1, b_2, \cdots, b_5$ are the estimates of the parameters $\alpha, \beta_1, \beta_2, \cdots, \beta_5$ of the logit model (1). More complicated logit models involving two-factor or higher-order effects need to be considered when model (1) does not fit the data.

After the model is chosen, its adequacy needs to be assessed. Procedures which have

been used thus far are rather informal and the statistical properties of measures of fit have not been studied. One of the most commonly used test of fit is the Pearson chi-square which is given by

$$\chi^2_P = \sum (o_i - m_i)^2 / m_i$$

where o is the observed frequency and m is the expected frequency. For smplicity the subscripts of i, j, k, l aind m are represented by a single subscript. On the other hand, Berkson (1946) proposed a least square like test statistic, called the logit chi-square. The logit chi-square is given by

$$\chi^2_{\text{Logit}} = \sum n_i p_i q_i (l_i - \hat{l}_i)^2, \text{ where } \hat{l}_i = \ln (\hat{p}_i / \hat{q}_i), \tag{3}$$

and is asymptotically distributed as chi-square, as is the Pearson chi-square. Here, p_i is the observed proportion of death among n_i infants and $\hat{p}_i$ is the estimated probability of death. And, l_i and $\hat{l}_i$ represents, respectively, observed and estimated value of the logit. Another test is the likelihood ratio chi-square test. This is less familiar than the Pearson chi-square, and is given by

$$\chi^2_{LR} = -2 \sum o_i \ln (m_i / o_i).$$

3. Comparison of Estimates

When logit models are concerned, three estimates stand out as possible candidates; minimum chi-square, minimum logit chi-square and maximum likelihood estimates. In many situations, the minimum chi-square and maximum likelihood estimates are identical, but they are not for the logit models. The normal equations to be solved for each estimate are given in the Appendix. Minimum chi-square and maximum likelihood estimates are efficient in the sense of Fisher and best asymptotically normal (BAN) in the sense of Neyman(1949). And, the minimum logit chi-square estimates are also, as shown by Taylor (1953), BAN estimates.

Quite often asymptotic optimality properties of the estimates and test statistics do not satisfy statisticians who are faced with the difficult problem of choosing the better one. Berkson(1968) compared minimum logit chi-square and maximum likelihood estimates in an analysis of data being reproduced in Table 2. In the table previous infant losses of mothers whose child presented behavioral problems are contrasted with losses of mothers of a comparable group of control children. The estimates corresponding to the observations of the individual cells by maximum likelihood and by minimum logit chi-

square are presented in the right-most columuns of Table 2 and these estimates are indeed very close. However, when the two estimates for the infant mortality data+ are compared, they are not quite close as Berkson stated. The estimates are calculated for the case of model (1) and are presented in the right-most columns of Table 1. Same phenomenon is observed when the estimates are calculated for the model that includes all possible two-factor interactions. Goodness of fit measures of chi-square test statistics also differ. For the model (1) the logit chi-square statistic is 22.39 on 26 degrees of freedom ($p=0.67$) and the likelihood ratio chi-square statistic is 29.04($p=0.31$). According to both test statistics it can be concluded that the model without interactions fits. However, the likelihood ratio chi-square statistic would not lead to a definite conclusion that the model with interactions fits ($\chi^2=24.71$, $p=0.01$), while the logit chi-square statistic shows that it does fit ($\chi^2=17.62, p=0.35$). This model with interactions is formulated by adding all possible two factor effects to the model (1).

If the mortality rate is not quite close to zero (or one), several methods of estimation may yield similar results. However, outside these limits estimates can diverge substantially. When Berkson compared two methods, the probability of response were ranging mostly from 0.20 to 0.80. In the analysis of infant mortality, rates do not exceed 0.10. Estimated parameters in a predictive model (1) of infant mortality are presented in Table 3 and the difference between estimated probability of death and observed probability of death are presented in Figure 1. Parameter estimation by maximum likelihood and by minimum logit chi-square method provides quite similar plots and the plot based on maximum likelihood estimates is shown in Figure 1. It shows that ill-fitting observations are those points especially close to zero.

It is shown that likelihood ratio chi-square statistic is sensitive to outlying responses, although in ideal situations they are known to have good optimality propesties. But how can one assess the impact of outlying observations on the various aspects of the fit? For instance, the assessment of the impact on parameter estimates and on fitted values of goodness-of-fit measures is the problem to be solved. Residuals are useful for the detec-

+ Logit analysis would be difficult without use of a computer. SAS(Statistical Analysis System) and BMDP (Biomedical Computer Program, P-series) are available for logit analysis. Also, there is a program, LINCAT, introduced by Forthofer, Starmer and Grizzle(1971). LINCAT Program provides the minimum logit chi-square for estimation and the logit chi-square for the goodness-of-fit test. However, SAS provides the maximum likelihood estimates and likelihood ratio chi-square in addition to the result of LINCAT program.

tion of ill-fitting observations, but they are not helpful for assessing their impact. A study in this direction is urged.

4. Selection of a Model

When the logit model is used in an attempt to estimate the impact of various predictors on infant mortality, selection of the best fitting model to the data is a major problem. Complicated models most often fit data more closely than a simpler model that is just a special case of the complicated one. On the other hand, a simpler model is often preferred over a more complicated one for the benefit of ease in interpretation. Thus, one needs to seek a method to aid in the selection of models. Unfortunately, however, there is not an all-purpose best method. Different approaches of model selection in loglinear models were suggested by Bishop(1969), Fienberg(1970), Goodman(1971) and Ku and Kullback(1968). And a stepwise procedure was suggested by Peduzzi et al. (1980).

In this section a relatively simple method of restricted chi-square test is used in conjunction with logit chi-square for the selection of a model. Neyman(1946) introduced a restricted chi-square test and showed that, with several different methods of producing best asymptotically normal(BAN) estimates, the restricted test criterion has asymptotically chi-square distribution as the sample size gets to infinity. A procedure of model selection is illustrated in Table 4, in which steps of selection can be traced starting at the top. First of all the simplest logit model can be constructed with a single variable and the appropriateness of the model itself against the most general alternative is tested by logit chi-square. Should the probability of getting larger than observed logit chi-square be small, the validity of the model is suspected and a more complicated model is sought. Since the model with a single variable of Educ does not fit the data too well ($\chi^2=37.94$, $p=0.15$), it would be natural to include more variables, so the next model tried is model (b) in Table 4. In this stage the added variable of birth interval is tested by the restricted chisquare test which is the difference of chi-square between models (b) and (a). Heuristically, this difference measures the increase in chi-square due to the additional restriction imposed by model (b) over those already imposed by model (a). The significance of added variables is tested by the differences of chi-squares. In the table model (e) includes simple terms of five predictor variables and model (f) all possible two-factor interaction terms plus simple terms. It is shown that the differences between models (b)

and (a) and between (d) and (b) are significant at the 5% level. However, the differences between models (e) and (d) and also between models (f) and (e) are not significant and hence a model with interaction terms is not suggested. One can settle that model (d) is the best model. It includes the variables of maternal education, birth interval and family's socioeconomic status.

5. Conclusion

Some variables examined in the analysis are shown to be affecting infant mortality. Simultaneously examining them, especially through the use of logit model, has clarified that variables of maternal education, birth interval and family's socioeconomic status have direct influence on mortality. Inclusion of interaction terms was investigated but they were omitted entirely due to their insignificance. However, one needs to be cautious since a few outlying responses may result unstable estimates and test statistics. The present analysis has been based on single births and thus familial effects for the infants born to the identical mother was not taken into consideration. Such information was not available to the present data and one needs to consider the possible effects in future studies.

As a result of viewing mortality as a response variable and the rest variables as predictors, the five-dimensional totals were conditioned, although they were not fixed by designing. When the distinction between the response and other predictor variables is not so important in the analysis, the loglinear model may well yield better estimates and this approach would not require conditioning on these totals.

APPENDIX

The normal equations to be solved for minimum chi-square, minimum logit chi-square and maximum likelihood estimates are given for a model with a single predictor variable. The minimum chi-square estimates are obtained by minimizing the Pearson chi-square, and the equations to be solved are shown to be

$$\sum n_i \frac{(\hat{p}_i q_i + \hat{q}_i p_i)}{\hat{p}_i \hat{q}_i}(p_i - \hat{p}_i) = 0,$$

$$\sum n_i \frac{(\hat{p}_i q_i + \hat{q}_i p_i)}{\hat{p}_i \hat{q}_i} x_i (p_i - \hat{p}_i) = 0,$$

where p_i is the observed proportion of death among n_i infants, $\hat{p}_i$ the estimated probability of death and x_i the value of predictor variable. As shown in equation (2), $\hat{p}_i$ is not linear in the parameters and neither the simultaneous equations are. Hence, iterative methods are required in solving. Secondly, from the likelihood equations of product multinomials, the equations to be solved for the maximum likelihood estimates are shown to be

$$\sum n_i(p_i-\hat{p}_i)=0,$$
$$\sum n_i x_i(p_i-\hat{p}_i)=0.$$

Again, iterative methods are required for the solution, since p_i is not linear in the parameters. Thirdly, the minimum logit chi-square estimates, suggested by Berkson (1946), are obtained by minimizing the logit chi-square of (3). The equations to be solved for the minimum logit chi-square estimates are shown to be

$$\sum n_i p_i q_i(l_i-\hat{l}_i)=0$$
$$\sum n_i p_i q_i x_i(l_i-\hat{l}_i)=0,$$

and they are solved explicitly since $\hat{l}_i$ is linear in the parameters as shown in (1).

REFERENCES

(1) Berkson, J. (1946). Approximation of Chi-square by Probits and by Logit, *J. Am. Stat. Ass.*, Vol. 41, 70—74.

(2) Berkson, J. (1968). Application of Minimum Logit χ^2 Estimate to a Problem of Grizzle with a Notation on the Problem of 'No Interaction', *Biometrics*, Vol. 24, 75—95.

(3) Bishop, Y. (1969). Full Contingency Tables, Logits and Split Continegncy Tables, *Biometrics*, Vol. 25, 383—400.

(4) Bross, D. and Shapiro, S. (1982). Direct and Indirect Associations of Five Factors with Infant Mortality, *Am. J. Epidem.*, Vol. 115, 78—91.

(5) Feldstein, M. (1966). A Binary Variable Multiple Regression Method of Analyzing Factors Affecting Peri-natal Mortality and Other Outcomes of Pregnancy, *J. Roy. Stat. Soc.*, Series A, Vol. 129, 61—73.

(6) Fienberg, S. (1970). The Analysis of Multidimensional Contingency Tables, *Ecology*, Vol. 51, 419—433.

(7) Forthofer, R.N., Starmer, C.F. and Grizzle, J. (1971). A Program for the Analysis of Categorical Data by Linear Models, *J. Biomed. System*, Vol. 2, 3—48.

(8) Goodman, L. (1971). The Analysis of Multidimensional Contingency Tables: Stepwise Procedure and Direct Estimation Methods for Building Models for Multiple Classifications, *Technometrics*, Vol. 13, 33—61.

(9) Ku, H. and Kullback, S. (1968). Interaction in Multidimensional Contingency Tables: An

Information Theoretic Approach, *J. Res. Nat. But. Standards*, Vol. 72, 159—199.

(10) Neutra, R.R., Fienberg, S.E., Greenland, S. and Friedman, E.A. (1978). Effects of Fetal Monitoring on Neonatal Death Rates, *New Eng. J. Med.*, Vol. 299, 324—326.

(11) Neyman, J. (1949). Contribution to the Theory of the χ^2 Test, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Univ. of California Press, 239—273.

(12) Peduzzi, P.N., Hardy, R.J. and Holford, T.R. (1980). A Stepwise Variable Selection Procedure for Nonlinear Regression Models, *Biometrics*, Vol. 36, 511—516.

(13) Quick, J.D., Greenlick, M.R. and Roghmann, K.J. (1981). Prenatal Care and Pregnancy Outcome in an HMO and General Population: A Multivariate Cohort Analysis, *Am. J. Public Health*, Vol. 71, 381—389.

(14) SAS Institute Inc., Raleigh (1979).

(15) Seo, K. (1982). *Usefulness of Mortality Rate as a Health Status Indicator*, PhD. Dissertation, Dept. of Medical Science, Yonsei University.

(16) Taylor, W. (1953). Distance Functions and Regular Best Asymptotically Normal Estimates, *Ann. Math. Stat.*, Vol. 24, 85—92.

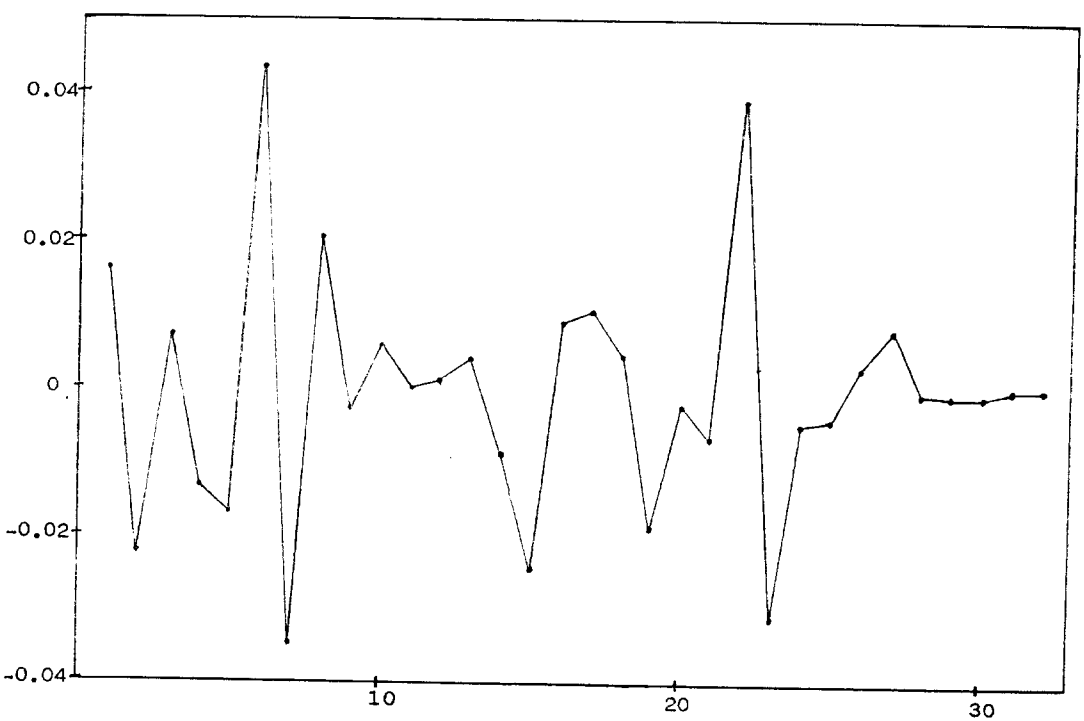

Fig 1. Residual plot based on the fitted logit model (1)

Table 1. Infant mortality data and estimated frequencies

Infant's sex	Maternal age	Maternal education	Survived	Observed: Low economic class, Birth interval: Under 2 yrs or 2 yrs over		Observed: High economic class, Birth interval: Under 2 yrs or 2 yrs over		Estimated: Maximum likelihood: Low economic class, Birth interval: Under 2 yrs or 2 yrs over		Estimated: Maximum likelihood: High economic class, Birth interval: Under 2 yrs or 2 yrs over		Estimated: Minimum logit chi-square: Low economic class, Birth interval: Under 2 yrs or 2 yrs over		Estimated: Minimum logit chi-square: High economic class, Birth interval: Under 2 yrs or 2 yrs over	
Male	20—29	Grammar	No	15	27	3	21	11.22	29.47	7.36	17.37	12.37	29.38	8.24	17.58
			Yes	220	815	197	625	223.78	812.53	192.64	628.63	222.63	812.62	191.76	628.42
		Some HS	No	4	8	4	22	3.27	7.78	8.35	20.75	3.75	8.05	9.74	21.84
			Yes	96	318	329	1114	96.73	318.22	324.65	1115.25	96.25	317.95	323.26	1114.16
	Under 20, 30 or over	Grammar	No	3	26	6	7	4.42	23.43	2.86	11.10	5.12	24.58	3.37	11.83
			Yes	84	602	67	380	82.58	604.57	70.14	375.90	81.88	603.42	69.63	375.17
		Some HS	No	0*	0*	3	8	.52	1.96	1.68	5.48	.63	2.14	2.07	6.07
			Yes	15	77	60	273	14.48	75.04	61.32	275.52	14.37	74.86	60.93	274.93
Female	20—29	Grammar	No	12	19	6	16	9.62	22.87	5.19	14.32	10.31	22.14	5.65	14.07
			Yes	212	708	151	577	214.58	704.13	151.81	578.68	213.69	704.86	151.35	578.93
		Some HS	No	1	12	6	17	2.82	8.59	6.60	17.44	3.14	8.46	7.47	17.82
			Yes	95	389	287	1047	93.18	392.41	286.40	1046.56	92.86	392.36	285.53	1046.18
	Under 20, 30 or over	Grammar	No	4	19	4	7	4.67	19.39	1.90	9.81	5.26	19.76	2.18	10.16
			Yes	98	559	50	374	97.33	558.61	52.10	371.19	96.74	558.24	51.82	370.84
		Some HS	No	0*	2	1	5	.28	1.74	1.27	4.52	.33	1.84	1.52	4.86
			Yes	8	74	52	253	7.72	74.26	51.73	253.48	7.67	74.16	51.48	253.14

* It is reported that adding small amounts to each cell, for example 0.5, may bias the variance estimates for problems where the total number of cell is large(SAS, 1979). Programs in SAS proceeds to take $0.5/n_{ijklm}$ if it meets a zero value.

Table 2. Berkson's (1968) data on the number of mothers with previous infant losses.

Birth order		Observed		Estimated			
				Maximum likelihood		Minimum logit chi-square	
		No. of mothers with		No. of mothers with		No. of mothers with	
		infant losses	no infant losses	infant losses	no infant losses	infant losses	no infant losses
2	Problems	20	82	20.503	81.497	20.500	81.500
	Controls	10	54	9.497	54.503	9.515	54.485
3—4	Problems	26	41	27.213	39.787	27.218	39.782
	Controls	16	30	14.787	31.213	14.814	31.186
5+	Problems	27	22	25.284	23.716	25.287	23.713
	Controls	14	23	15.716	21.284	15.740	21.260

Table 3. Maximum likelihood and minimum logit chi-suare estimates of the coefficients of model (1)

MODEL(1)		
Variables in the model	Minimum logit chi-square estimates	Maximum likelihood estimates
Sex	0.0703	0.0551
Birth interval	0.2148	0.1617
Age	−0.0597	−0.0331
Educ	0.1778	0.1976
Socioeco status	0.1282	0.1361

Table 4. Restricted chi square values for logit models applied to the infant mortality data

Variables included in the model	χ^2	d.f.	difference of χ^2	difference of d.f.
Sex	51.45++	30		
Age	50.72+	30		
Socioeco status	42.84	30		
Birth interval	42.26	30		
(a) Educ	37.94	30		
(b) Educ, birth interval	28.79	29		
Difference between models (b) and (a)			9.15++	1
(c) Educ,birth interval, educ* birth interval	26.77	28		
Difference between models (c) and (b)			2.02	1
(d) Educ, birth interval, socioeco status	24.65	28		
Difference between models (d) and (b)			4.14+	1
Educ, birth interval, socioeco status	22.64	27		
educ* birth interval			2.01	1
Educ, birth interval, socioeco status	24.38	27		

educ* socioeco status			.26	1
Educ, birth interval, socioeco status	24.60	27		
birth interval* socioeco status			.04	1
(e) Sex, age, socioeco status,	22.39	26		
birth interval, educ			2.26	2
Difference between models (e) and (d)				
(f) Sex, age, socioeco status, educ,	17.62	16		
birth interval, sex* age,				
sex* socioeco status, sex* educ,				
sex* birth interval,				
age* socioeco status, age* educ,				
age* birth interval,				
socioeco status* educ,				
socioeco status* birth interval,				
educ* birth interval				
Difference between models (f) and (e)			4.77	10

+(or ++) indicates that value is in upper 5% (or 1%) tail of the corresponding χ^2 distribution with d.f. as indicated.