Log-linear Modelling with Complex Survey Data

Chris Skinner
University of Southampton
United Kingdom

Abstract: Different approaches have been proposed for fitting log linear models to contingency tables based on complex survey data. One approach seeks to handle differential sampling via an offset term in the model. A second pseudo maximum likelihood approach fits models to the weighted table. This paper will investigate and compare the properties of inference procedures for these methods. Point estimation, variance estimation and testing will be considered. The case when survey weights are constant within the cells of the table will be contrasted with the case when they are variable. Some empirical illustration will be provided using data from a survey in France collecting data on daughter's and father's social class.

1. Introduction

Categorical variables are common in social survey data. A natural approach to analysing data on categorical variables is via Poisson log linear models. This paper considers the question of how to take account of complex sampling schemes when fitting such models. For more background and literature references see Skinner and Vallet (2010).

2. The Log-linear Model and Bernoulli Sampling

Consider a contingency table formed by cross-classifying two or more categorical variables. Let k denote a generic cell in the table ($k = 1,...,M$) and let N_k denote the population frequency in cell k. Suppose the N_k are generated as outcomes of independent Poisson random variables with means μ_k. We consider the log linear model:

$$\log(\mu) = X\lambda , \tag{1}$$

where $\log(\mu)$ is the $M \times 1$ vector with elements $\log(\mu_k)$, X is an $M \times p$ model matrix containing specified values, usually either 0 or 1, and λ is a $p \times 1$ vector of unknown parameters, where $p \leq M$.

The Poisson distribution has the property that it is preserved under Bernoulli sampling. Thus, if Bernoulli sampling with inclusion probability π_k is used in cell k then the sample frequency n_k in this cell is also Poisson distributed with mean $\pi_k \mu_k$. Hence, a log-linear model for the sample frequencies, may be expressed as:

$$\log(\mu_s) = \log(\pi) + X\lambda, \qquad (2)$$

where $\log(\mu_s)$ and $\log(\pi)$ are the $M \times 1$ vectors containing $\log(\pi_k \mu_k)$ and $\log(\pi_k)$ respectively. The parameter vector λ may then be estimated using maximum likelihood (ML), by treating $\log(\pi)$ as an offset term in the model.

3. Estimation under More General Sampling Schemes

The use of ML to estimate λ above depends on the restrictive assumption that the inclusion probabilities are fixed within cells. We now consider the case when these probabilities may vary between units (and, in particular, within cells).

Consider, for simplicity, the case when there is only a finite number of possible values of the sample inclusion probabilities, denoted $\pi_1, \pi_2, ..., \pi_H$. We shall refer to the different parts of the population which are sampled with different probabilities as strata and assume that units in stratum *h* are selected by Bernoulli sampling with probability π_h ($h = 1, 2, ..., H$).

Let N_{kh} be the population count in cell *k* in stratum *h*, so that $N_k = \sum_{h=1}^{H} N_{kh}$. Suppose that the N_{kh} are generated independently as Poisson random variables: $N_{kh} \sim Poisson(\mu_{kh})$. This implies that $N_k \sim Poisson(\mu_k)$, where $\mu_k = \sum_{h=1}^{H} \mu_{kh}$, and also that the numbers n_{kh} of sample units which fall into cell *k* and stratum *h* are independently distributed as: $n_{kh} \sim Poisson(\pi_h \mu_{kh})$. It follows that the distribution of $n_k = \sum_{h=1}^{H} n_{kh}$ is also Poisson i.e. $n_k \sim Poisson(\mu_{sk})$, where $\mu_{sk} = \sum_{h=1}^{H} \pi_h \mu_{kh}$.

These assumptions thus preserve the assumption in section 2 that the sample frequencies in the cells are Poisson distributed but allow for possible variation in the inclusion probabilities. Assuming that the population frequencies obey the log-linear model in (1), it no longer appears possible to express the model for the sample frequencies as in (2), where the inclusion probabilities only appear as an offset term.

A full ML approach for the stratified sampling approach above would seem to require a specification of a model for the μ_{kh}. We shall not pursue such an approach here, assuming that the analyst only wishes to specify a model for the population, aggregated across strata.

An alternative simpler approach was proposed by Clogg and Eliason (1987). They viewed π_k as the expected sampling fraction n_k / N_k in cell *k*, and proposed to estimate this fraction by by $z_k = n_k / \hat{N}_k$, where $\hat{N}_k$ is the sum of survey weights across sample units in cell *k*, i.e. the Horvitz-Thompson estimator of N_k. They then modified model (2) to the following expression:

$$\log(\mu_s) = \log(z) + X\lambda \tag{3}$$

where $\log(z)$ is the $M \times 1$ vector containing the $\log(z_k)$ and proposed to fit this model using conventional ML methods, treating $\log(z)$ as an offset.

Expression (3) does not strictly provide a model for the expected values of the sample frequencies, μ_s, since $\log(z)$ is sample-dependent. If we replace n_k in z_k by its expectation, the kth element of μ_s, we see that these terms cancel in (3) and this expression may be rewritten as:

$$\log(\hat{N}) = X\lambda \tag{4}$$

where $\log(\hat{N})$ is the $M \times 1$ vector containing the $\log(\hat{N}_k)$. This corresponds to a pseudo ML approach (Rao and Thomas, 1988). In fact, Skinner and Vallet (2010) show that applying conventional ML fitting to (3) does not lead to an identical point estimator to the pseudo ML estimator. Both estimators may be expressed as solutions of the estimating equation:

$$\sum_k a_k \{\hat{N}_k - \exp(x_k \lambda)\} x_k = 0, \tag{5}$$

where x_k is the kth row of X, $a_k = 1$ for the pseudo ML approach and $a_k = z_k$ for the Clogg-Eliason approach. Both estimators are consistent when model (1) holds although they may converge to different values if the model fails. Considering the class of estimators defined by alternative values of a_k and assuming that model (1) and the stratified sampling assumptions above hold, the Clogg-Eliason point estimator has minimum large sample variance if the stratum selection probabilities are constant within cells. In general, however, this approach will not be optimally efficient. The pseudo ML point estimator will be efficient when the stratification variable is essentially independent of the variable defining the cells, but not in general.

4. Variance Estimation

Skinner and Vallet (2010) show that the use of conventional ML estimation methods based upon expression (3), treating $\log(z)$ as an offset as proposed by Clogg and Eliason (1987), leads in general to inconsistent variance estimation and, in particular, to underestimation of standard errors. They propose that, instead, conventional linearization or replication methods are employed for variance estimation.

5. Empirical Comparison of Approaches

We now set out to compare the Clogg and Eliason (1987) and pseudo ML (hereafter denoted CE and PML) approaches empirically. Further details are provided in Skinner and Vallet (2010). We use data from the *1985* Enquête Formation & Qualification Professionnelle, a survey with complex sampling design conducted by the French Statistical Office. The survey covered persons aged between 13 and 69 in 1982. It was administered to a stratified sample of 46,500 individuals drawn from a 1982 master sample with stratum sampling fractions that varied between about 1/200 and 1/2690. There were 73 strata, defined by nationality, labour market position, socio-economic class and age group. Weights were constructed to take account not only of the different sampling fractions but also for missing data. The weights were ratios of census counts to counts of survey respondents within weighting classes defined by the strata cross-classified with residential area at the census (rural, urban, or Parisian).

The analysis here is restricted to the sub-sample of 5,159 women, with French nationality at the date of the survey, aged between 35 and 59 at the end of December 1985, currently employed at the date of the survey, and who reported information about their current socio-economic class and their father's socio-economic class when they stopped attending school or university on a regular basis. The 5,159 women belong to 18 different strata with initial sampling fractions varying between 1/310 and 1/2500.

The analysis is based on the 7 x 7 two-way contingency table that cross-classifies women's socio-economic class with their father's socio-economic class when they stopped attending school or university on a regular basis. The mobility table uses a discrete and unordered socio-economic classification defined as follows: (1) higher-grade salaried professionals; (2) company managers and liberal professions; (3) lower-grade salaried professionals; (4) artisans and shopkeepers; (5) non-manual workers; (6) foremen and manual workers; (7) farmers. Table 1 presents both unweighted frequencies and weighted frequencies in the mobility table after rescaling the latter to the exact sample size.

The objective of the analysis is to investigate the structure and strength of the association between father's socio-economic class and daughter's socio-economic class in 1985 within French society. A log-linear model proposed by Hauser (1978, 1980) is employed, which defines two-way interaction effects by constraining some of them to be equal across cells of the contingency table. Assuming that i and j respectively index father's class and daughter's class, that the cells (i, j) are assigned to K mutually exclusive and exhaustive subsets and that each of those sets shares a common interaction parameter δ_k, the logged expected frequency in cell (i, j) of the mobility table is expressed as follows:

$$log\, \mu_{ij} = \alpha + \beta_i + \gamma_j + \delta_k \qquad \text{if the cell } (i, j) \text{ belongs to subset } k.$$

Thus, aside from total (α), row (β_i), and column (γ_j) effects, each expected frequency is determined by only one interaction parameter (δ_k) which "reflects the density of mobility or immobility in that cell relative to that in other cells in the table" (Hauser, 1980, p.416). The interaction parameters of the model may therefore "be interpreted as indexes of the social distance between categories of the row and column classifications" (Hauser, 1980, p.416).

Vallet (2005) used sociological hypotheses to build such a model of the father-daughter mobility table with $K = 7$ interaction parameters. The specification of the subsets of cells in the initial and final models presented in Table 3 is discussed in Skinner and Vallet (2010). Goodness of fit tests presented by Mason (2010) suggest that the final model provides a good fit to the data. The initial model fits less well and goodness of fit tests are on the verge of rejecting it at the 95% level.

For the initial and final models, estimates and standard errors are now presented based upon four different approaches: the standard ML approach for the tables of unweighted frequencies and of weighted rescaled frequencies; the CE approach; and the PML approach. For details of computation, see Skinner and Vallet (2010).

Table 3 presents parameter estimates and standard errors obtained for the initial and final models under all four approaches. Consider the point estimates first. The estimates obtained by applying the standard ML approach to the weighted rescaled table are identical to those from the PML approach as expected. Thus, there are really just three sets of point estimates to compare. The most marked differences are between the unweighted estimates and the other two (PML and CE) estimates. As discussed earlier, both these estimators will be approximately unbiased if the model is true. We cannot be certain that either of the models is true but it seems reasonable to view the differences between the unweighted estimates and the other two estimates as evidence of bias in the former procedure. This bias is especially pronounced in the case of the γ_j parameters and this may be attributed to the strong correlation between the column variable (women's socio-economic class in 1985) and one of the stratifying variables (women's socio-economic class at the census) upon which the sampling is differential. The PML and CE estimates are broadly similar and should not lead to any difference in substantive interpretation for either model. Leaving aside consideration of the standard errors, there seems no strong reason to prefer one set of estimates to the other. One possible argument in favour of the PML estimator, following Patterson et al. (2002) and mentioned in section 6.2, is that the PML estimator is 'estimating' a well-defined population quantity if the model is false, whereas the CE estimator is then estimating a quantity dependent on the arbitrariness of the sampling scheme.

As regards standard errors in Table 3, only those for the PML estimator have been estimated in a way which takes appropriate account of the complex sampling. Since the weighted rescaled and the PML point estimators are identical, the differences between the standard errors for these two estimators demonstrate that the former method can often lead to seriously incorrect standard errors. Standard errors were also calculated for the unweighted point estimators using the jackknife method and that these too can differ from the values in Table 3, although the differences are more minor. No further comment is made on these results, however, since the unweighted point estimators show clear bias and their standard errors are of little interest.

The standard errors of the CE point estimator obtained via a valid jackknife approach are compared in Table 4 with those obtained via the CE approach. The CE approach is seen to underestimate the standard errors uniformly. The jackknife value is often at least 10% higher and sometimes at least 20% higher. This empirical investigation therefore illustrates how the CE variance estimator can systematically underestimate the true variability. Moreover, in Table 3 the standard errors obtained under the CE approach are virtually identical to those of the unweighted approach. Hence the device of including the offset term in the model seems to provide virtually no benefit in capturing the effect of unequal sampling weights on the standard error.

Finally, the jackknife estimates for the CE estimator in Table 4 may be compared with the jackknife estimates for the PML estimator in Table 3. These are very similar. This is not surprising since the values of the point estimators were similar too. It implies that, at least for this application, there is no evidence of an efficiency advantage of the CE point estimator compared to the PML approach.

6. Conclusions

Approaches to estimating a log linear model in the presence of complex sampling schemes have been compared, both theoretically and via an empirical study. One approach, proposed by Clogg and Eliason (1987), adds an offset term to the model and employs conventional maximum likelihood. Another approach is pseudo ML. With respect to point estimation, little reason was found to prefer one method over the other. With respect to variance estimation, the use of the standard ML variance estimation approach proposed by Clogg and Eliason was found to lead to underestimation and is not recommended. The pseudo ML approach provided appropriate standard errors.

Acknowledgements

The empirical study in section 5 was conducted by Louis-André Vallet. Thanks are due to Natalie Shlomo for programming the jackknife variance estimates in Table 4.

References

Clogg, C.C. and S.R. Eliason (1987) Some common problems in log-linear analysis. *Sociological Methods & Research* 16:8-44.

Hauser, R.M. (1978) A structural model of the mobility Table. *Social Forces* 56:919-53.

Hauser, R.M. (1980) Some exploratory methods for modeling mobility tables and other cross-classified data. *Sociological Methodology* 11:413-58.

Mason, B. (2010) A comparison of log-linear modelling approaches for complex survey data. M.Sc. Social Statistics dissertation, University of Southampton.

Patterson, B.H., C. Mitchell Dayton, and B.I. Graubard (2002) Latent class analysis of complex sample survey data: application to dietary data. *Journal of the American Statistical Association* 97:721-29.

Rao, J. N. K. and D. R. Thomas (1988) The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology* 18:213-69.

Skinner, C. and Vallet, L.-A. (2010) Fitting log-linear models to contingency tables from surveys with complex sampling designs: an investigation of the Clogg-Eliason approach. *Sociological Methods & Research*, 39:83-108.

Vallet, L.-A. (2005) Utiliser le modèle log-linéaire pour mettre au jour la structure du lien entre les deux variables d'un tableau de contingence : un exemple d'application à la mobilité sociale. *Actes des Journées de Méthodologie Statistique 2005*, Paris, Insee, 21 p.

Table 1 – Unweighted frequencies and weighted (rescaled) frequencies in the mobility table

Father's class \ Daughter's class	Frequency	1	2	3	4	5	6	7	Total
1 Higher-grade salaried professionals	Unweighted	164.00	25.00	136.00	12.00	59.00	9.00	0.00	405.00
	Weighted	81.23	13.01	113.18	15.35	66.32	8.08	0.00	297.17
2 Company managers and liberal professions	Unweighted	56.00	27.00	37.00	14.00	28.00	3.00	3.00	168.00
	Weighted	28.78	11.72	38.22	14.46	32.45	2.65	7.01	135.29
3 Lower-grade salaried professionals	Unweighted	95.00	16.00	161.00	15.00	115.00	18.00	4.00	424.00
	Weighted	48.08	11.44	129.70	22.79	131.79	18.20	4.77	366.78
4 Artisans and shopkeepers	Unweighted	97.00	35.00	219.00	78.00	200.00	35.00	8.00	672.00
	Weighted	52.25	21.35	174.45	118.41	223.37	39.57	14.27	643.67
5 Non-manual workers	Unweighted	59.00	7.00	145.00	32.00	182.00	29.00	3.00	457.00
	Weighted	30.18	3.68	120.03	53.42	216.57	28.65	4.17	456.70
6 Foremen and manual workers	Unweighted	128.00	18.00	419.00	124.00	930.00	339.00	37.00	1995.00
	Weighted	64.18	14.88	361.46	184.12	1065.19	355.76	47.06	2092.66
7 Farmers	Unweighted	38.00	8.00	164.00	73.00	342.00	136.00	277.00	1038.00
	Weighted	20.29	5.63	134.71	101.98	394.83	140.49	368.80	1166.73
Total	Unweighted	637.00	136.00	1281.00	348.00	1856.00	569.00	332.00	5159.00
	Weighted	324.99	81.71	1071.75	510.54	2130.52	593.40	446.08	5159.00

Note: Weighted frequencies are rescaled to the sample size by multiplying them by the ratio 5159/4386881.

Table 2 – Initial model and final model for the structure of the association in the mobility table

Initial model	1	2	3	4	5	6	7
1 – Higher-grade salaried professionals	II	III	IV	V	VI	VII	VII
2 – Company managers and liberal professions	III	II	IV	IV	VI	VII	VII
3 – Lower-grade salaried professionals	IV	IV	IV	V	V	VI	VII
4 – Artisans and shopkeepers	V	IV	V	IV	V	VI	VI
5 – Non-manual workers	VI	VI	V	V	V	V	VI
6 – Foremen and manual workers	VII	VII	VI	VI	V	IV	V
7 – Farmers	VII	VII	VII	VI	VI	V	I

Final model	1	2	3	4	5	6	7
1 – Higher-grade salaried professionals	II	II	III	IV	V	VI	VII
2 – Company managers and liberal professions	II	II	III	III	V	VI	IV
3 – Lower-grade salaried professionals	III	III	III	IV	IV	V	VI
4 – Artisans and shopkeepers	IV	III	IV	III	V	V	V
5 – Non-manual workers	V	V	IV	IV	IV	V	V
6 – Foremen and manual workers	VI	VI	V	IV	IV	III	IV
7 – Farmers	VII	VI	VI	IV	V	IV	I

Note: Rows and columns in the matrices respectively correspond to father's socio-economic class and daughter's socio-economic class. Among the interaction effects, I is supposed to be the strongest and VII the weakest.

Table 3 – Comparison of parameter estimates and standard errors (in parentheses)

Parameter	Initial model				Final model			
	Unweighted	Weighted rescaled	Clogg & Eliason	Pseudo maximum likelihood	Unweighted	Weighted rescaled	Clogg & Eliason	Pseudo maximum likelihood
β_1 (se)	-1.813 (0.087)	-1.825 (0.086)	-1.828 (0.086)	-1.825 (0.098)	-1.747 (0.084)	-1.754 (0.083)	-1.763 (0.083)	-1.754 (0.093)
β_2 (se)	-2.626 (0.107)	-2.621 (0.108)	-2.612 (0.106)	-2.621 (0.133)	-2.663 (0.102)	-2.610 (0.105)	-2.632 (0.102)	-2.610 (0.125)
β_3 (se)	-1.532 (0.079)	-1.559 (0.078)	-1.549 (0.079)	-1.559 (0.090)	-1.492 (0.076)	-1.517 (0.075)	-1.514 (0.076)	-1.517 (0.085)
β_4 (se)	-0.856 (0.069)	-0.857 (0.067)	-0.855 (0.070)	-0.857 (0.079)	-0.633 (0.061)	-0.614 (0.059)	-0.643 (0.061)	-0.614 (0.068)
β_5 (se)	-1.134 (0.072)	-1.104 (0.072)	-1.111 (0.073)	-1.104 (0.082)	-1.036 (0.067)	-1.013 (0.065)	-1.021 (0.067)	-1.013 (0.075)
β_6 (se)	0.492 (0.049)	0.510 (0.049)	0.505 (0.049)	0.510 (0.056)	0.487 (0.048)	0.507 (0.047)	0.497 (0.048)	0.507 (0.054)
β_7	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0
γ_1 (se)	2.187 (0.149)	1.179 (0.139)	1.261 (0.149)	1.179 (0.166)	2.177 (0.148)	1.196 (0.138)	1.238 (0.148)	1.196 (0.157)
γ_2 (se)	0.585 (0.169)	-0.269 (0.169)	-0.182 (0.170)	-0.269 (0.205)	0.450 (0.167)	-0.373 (0.167)	-0.321 (0.167)	-0.373 (0.198)
γ_3 (se)	2.889 (0.140)	2.360 (0.120)	2.424 (0.140)	2.360 (0.150)	2.855 (0.139)	2.341 (0.119)	2.376 (0.139)	2.341 (0.146)
γ_4 (se)	1.473 (0.147)	1.508 (0.124)	1.555 (0.148)	1.508 (0.156)	1.204 (0.147)	1.253 (0.124)	1.282 (0.147)	1.253 (0.153)
γ_5 (se)	3.089 (0.137)	2.895 (0.116)	2.943 (0.137)	2.895 (0.143)	3.167 (0.137)	2.971 (0.116)	3.003 (0.137)	2.971 (0.144)
γ_6 (se)	1.605 (0.146)	1.297 (0.126)	1.349 (0.146)	1.297 (0.150)	1.638 (0.146)	1.340 (0.126)	1.370 (0.146)	1.340 (0.150)
γ_7	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0
δ_I (se)	3.561 (0.163)	3.451 (0.146)	3.569 (0.163)	3.451 (0.189)	4.163 (0.228)	4.096 (0.266)	4.138 (0.228)	4.096 (0.252)
δ_{II} (se)	2.730 (0.119)	2.619 (0.147)	2.660 (0.118)	2.619 (0.135)	3.215 (0.191)	3.104 (0.251)	3.123 (0.191)	3.104 (0.214)
δ_{III} (se)	2.396 (0.150)	2.297 (0.189)	2.326 (0.149)	2.297 (0.186)	2.276 (0.187)	2.252 (0.245)	2.275 (0.187)	2.252 (0.208)
δ_{IV} (se)	1.683 (0.086)	1.633 (0.093)	1.700 (0.085)	1.633 (0.105)	1.692 (0.183)	1.658 (0.243)	1.675 (0.183)	1.658 (0.204)
δ_V (se)	1.161 (0.084)	1.078 (0.092)	1.154 (0.084)	1.078 (0.103)	1.245 (0.181)	1.217 (0.241)	1.240 (0.181)	1.217 (0.201)
δ_{VI} (se)	0.683 (0.072)	0.641 (0.080)	0.699 (0.072)	0.641 (0.087)	0.731 (0.177)	0.708 (0.239)	0.702 (0.177)	0.708 (0.196)
δ_{VII}	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0
Deviance	86.11	77.12	75.58	-	47.71	33.69	34.77	-
DF	29	29	29	-	29	29	29	-

Table 4 – Comparison of estimated standard errors for Clogg-Eliason estimator: Clogg-Eliason approach vs Jackknife method allowing for complex design

Parameter	Initial model		Final model	
	Clogg & Eliason	Jackknife	Clogg & Eliason	Jackknife
β_1	0.086	0.102	0.083	0.096
β_2	0.106	0.130	0.102	0.122
β_3	0.079	0.090	0.076	0.086
β_4	0.070	0.078	0.061	0.068
β_5	0.073	0.081	0.067	0.074
β_6	0.049	0.055	0.048	0.054
β_7	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0
γ_1	0.149	0.158	0.148	0.155
γ_2	0.170	0.204	0.167	0.201
γ_3	0.140	0.144	0.139	0.143
γ_4	0.148	0.152	0.147	0.149
γ_5	0.137	0.140	0.137	0.141
γ_6	0.146	0.147	0.146	0.148
γ_7	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0
δ_I	0.163	0.181	0.228	0.253
δ_{II}	0.118	0.139	0.191	0.218
δ_{III}	0.149	0.192	0.187	0.212
δ_{IV}	0.085	0.101	0.183	0.207
δ_V	0.084	0.099	0.181	0.204
δ_{VI}	0.072	0.084	0.177	0.200
δ_{VII}	Fixed at 0	Fixed at 0	Fixed at 0	Fixed at 0