Latent class analysis of complex sample survey data

Vermunt, J.K.

Published in:
Journal of the American Statistical Association

Publication date:
2002

Comments on "Latent class analysis of complex sampling data" by Jeroen K. Vermunt, Tilburg University

Patterson, Dayton, and Graubard (PDG) show how to take into account complex sampling designs in latent class (LC) modeling. Sampling weights are dealt with by pseudo-maximum likelihood (PML) estimation, a method that was also used by Wedel, Ter Hofstede, and Steenkamp (1998) for mixture modeling and that is implemented in some LC software packages such as Latent GOLD (Vermunt and Magidson, 2000). Because standard asymptotic theory is no longer valid, PDG propose estimating standard errors by means of a simple but computationally intensive jackknife procedure that simultaneously corrects for stratification, clustering, and weighting.

In the discussion, I focus on the question of whether to use sampling weights in LC modeling, I advocate the linearization variance estimator, present a maximum likelihood (ML) estimator, propose a random-effects LC model, and give an alternative analysis of the dietary data which takes into account the longitudinal nature of the data.

Weighting – yes or no? I am not convinced that in the presented application the weighted solution is better than the unweighted solution. In order to clarify this point, it is important to make a distinction between the two types of parameters in the LC model; that is, the LC proportions θ_l and the item conditional probabilities α_{ljr}. It is clear that the unweighted estimates of θ_l will be biased if characteristics correlated with the sampling weights are also correlated with class membership. However, it is important to note that the results obtained with a standard LC analysis are only valid if the population is homogenous with respect to the α_{ljr}. If this assumption holds, there is no need to use sampling weights for the estimation of the α_{ljr}; and if it does not hold, use of sampling weights does not solve the

problem. Heterogeneity in α_{ljr} should be dealt with by introducing the relevant grouping variables in a multiple-group LC analysis.

Taking into account the much larger standard errors in the weighted analysis, I prefer the unweighted $\widehat{\alpha}_{ljr}$. Possible biases in the unweighted $\widehat{\theta}_l$ can be corrected by reestimating the LC probabilities, say by PML, fixing the α_{ljr} at their unweighted ML estimates. This two-step estimator yields an estimated LC proportion of .35, which is quite close to the unweighted estimate of .33. Such a small upwards correction of the number of low consumers is what could be expected from the fact that weighting increases the observed proportion of non-consumers. A weighted analysis with the PML method, however, yields a downwards correction of the proportion of low consumers ($\widehat{\theta}_1$=.18).

Linearization estimator Wedel, Ter Hofstede, and Steenkamp (1998) proposed using a linearization or robust variance estimator in mixture modeling with complex samples. The method is described in detail by Skinner et al. (1989: 83). PDG state that this approach is less flexible in that it requires developing new software. I do not agree with this statement as the method is easily implemented in any LC software that already computes first and second derivatives of the pseudo-likelihood function. It should be noted that contrary to PDG's jackknife method, the additional computation time is negligible.

The standard errors I obtained with the linearization estimator are very close to the jackknife standard errors. Actually, they are slightly smaller, which indicates that they are not only easier and faster to obtain, but also somewhat better given that PDG's simulation study showed that the jackknife slightly overestimates the standard errors.

ML estimation of LC models with sampling weights Clogg and Eliason (1987) and Magidson (1987) proposed a ML estimator for log-linear models with sampling weights under Poisson sampling. Let k denote a particular response pattern, and let δ_{ik} be 1 if case i has response pattern k and 0 otherwise. The unweighted frequency in cell k, n_k, equals $\sum_i \delta_{ik}$ and the weighted frequency, $n_k^{(w)}$, is obtained by $\sum_i \delta_{ik} w_i$. The inverse of the cell-specific sampling weight, z_k, equals $n_k/n_k^{(w)}$. The log-linear model that is used in a weighted analysis has the following form

$$m_k = \exp(\mathbf{x}_k\beta)\ z_k.$$

The term $\exp(\mathbf{x}_k\beta)$ defines an expected cell entry in the population, while the corresponding expected cell entry in the "biased population", m_k, is obtained by multiplying it by z_k.

Under Poisson sampling, ML estimation of the unknown β parameters involves maximizing $\log L = \sum_k [n_k \ln(m_k) - m_k]$. This function correctly reflects the data generating process as far as the unequal selection (or nonresponse) probabilities are concerned. Note that the PML method maximizes $\log PL = \sum_k \left[n_k^{(w)}(\mathbf{x}_k\beta) - \exp(\mathbf{x}_k\beta)\right]$, which is clearly not the same.

The above method can easily be generalized to LC models if we write the LC model as a log-linear model for an incomplete table. Using l as the index for the latent classes, the model for m_k is now

$$m_k = \left[\sum_l \exp(\mathbf{x}_{lk}\beta)\right]\ z_k,$$

where the linear term $\mathbf{x}_{lk}\beta$ defines the LC model (see Haberman, 1979). The Newton (Haberman, 1988) and LEM (Vermunt, 1997) programs for log-linear modeling with incomplete tables can be used to implement this method.

Application of this ML method to the dietary data yields results that are similar to the

PDG's PML results. An advantage is, however, that standard goodness of-fit measures can be used to assess model fit. The likelihood-ratio statistic L^2 equals 18.32 ($df = 6$ and $p = 0.01$), indicating that the 2-class model does not fit the data.

Random-effects latent models A standard method for dealing with clustering effects is random-effects modeling. In the application, a cluster is a PSU within a stratum, say PSU h in stratum s, denoted by sh. Let us assume that the LC proportions are coefficients that vary between PSU's. A simple random-effects two-class model is obtained by assuming that $\ln(\theta_{1(sh)}/\theta_{2(sh)}) \sim N(\mu, \sigma^2)$. The contribution of cluster sh to the log-likelihood function equals

$$\ln L_{sh} = \ln \int \left\{ \prod_{\text{all } i \text{ in cluster } sh} \left(\sum_{l=1}^{L} \theta_{l(sh)} P(\mathbf{Y}_i | c_l) \right) \right\} f(\theta_{(sh)} | \mu, \sigma^2) d\theta_{(sh)}.$$

The integral can, for instance, be solved by Gauss-Hermite quadrature.

Application of this random-effects LC model to the (unweighted) dietary data revealed that there is no evidence for variation of the LC proportions between clusters. This is in agreement PDG's results.

Measurement error or change? As indicated by PDG, the four dietary recalls were obtained at six time points; that is, recalls 2-4 do not represent the same recall occasions for all of the women. In order to be able to take the longitudinal nature of the data into account, I reanalyzed the (unweighted) data using six occasions instead of four, where each woman has two missing values. It should be noted that as long as the missing data can be assumed to be missing at random, it does not cause special problems within a ML framework.

First, I estimated standard LC models with different numbers of classes. The two-class model turned out to be the best in terms of fit (L^2=52.07, df=50, and p=0.39). Equating

all time-specific intake probabilities for the high-consumption class and the ones of the first three time points for the low-consumption class did not cause the fit to deteriorate (L^2=55.82, df=57, and p=0.52). The estimated intake probability was 0.80 for the high- and stable-consumption class. The low-consumption class had 0.57 at the first three time points, dropped to 0.38 and 0.20, and increased to 0.46 at the last time point.

PDG do not pay attention to the fact that there is not only measurement error in the reported intake, but also change in intake over time. The LC model, however, can not make a distinction between measurement error and change. A model that is better suited for this purpose is a hidden or latent Markov model. A simple hidden Markov with two latent states and time-invariant measurement errors fits almost as good as the two-class LC model (L^2=54.37, df=50, p =0.31), but tells a more interesting story about the same data set. The high-intake class has an intake probability of 0.83 at each time point and the low-intake class of 0.36. Note that these measurement errors (0.17 and 0.36) are smaller than in the standard LC model. Between occasions one and three there are similar numbers of moves from high to low intake as from low to high, between time points three and five there are much more moves from high to low, and between time points five and six there are much more moves from low to high. This indicates that besides measurement error there is a season effect in the consumption of vegetables: the proportion of low consumers depends on the period of the year.

References

Haberman, S.J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observations. *Sociological Methodology*, 18, 193-211.

Magidson, J. (1987). *Weighted log-linear modeling.* American Statistical Association, Proceedings of Social Statistics Section, 171-174.