Homogeneous Linear Predictor Models Specified by Constraints on the Goodman and Kruskal Tau Index

Homogeneous Linear Predictor Models Specified by Constraints on the Goodman and Kruskal Tau Index

Elena Siletti
Università di Milano – Bicocca, elena.siletti@unimi.it

Keywords: homogeneous linear predictor models, categorical data, measure of association, measure of variation, contingency tables.

1. Introduction

In this work we focus our research in Lang's HLP models for contingency tables, constraining the expected table counts through the Goodman and Kruskal tau-*b* and, or the Gini's index. Although we have used them, a lot of other measures can be used in this kind of models. For example, to measure association in discrete variable we could use odds ratios or coefficients derived from them such as the Goodman and Kruskal's gamma or Kendall's tau. For HLP models link function is allowed to be many-to-one and nonlinear. If the link functions are many to one as in our work, it is generally not possible to re-express the likelihood in terms of the linear predictor parameter β alone. This is one of the main reason that the ML approach has in the past typically been abandoned in favour of alternative fitting methods. Although less of an issue today, ML estimation was also avoided in the past because of computational complexity. Following Lang's approach we use ML estimation which is an attractive alternative to non iterative weighted least squares for several reason: including model based estimates of cell probabilities along with cell specific residuals are available; likelihood ratio and score statistics are available and, unlike the Wald statistics, are invariant to the model parameterization; profile likelihood confidence intervals are available; smoothed estimates of the link function variance are used, mitigating potential problems with zero counts; the ML estimation method does not require full rank link function Jacobian; ML estimates are invariant; and estimators have higher order efficiency properties that are not shared by WLS estimators.
In order to calculate the derivatives respect to the joint probabilities of the link functions we have introduced the exp-log notation, in this way, only the derivative matrix of this formula has been implemented and derived. Then, because following Lang the constraint functions for HLP models must be function of the single variable **m**, to use the Goodman and Kruskal measure of association and the Gini's heterogeneity index as constraints, we have specified them in the exp-log notation using the expected counts.
To use them in existing Lang's computer procedure mph.fit, we have written some new R's functions. Two functions which calculate the measures using expected counts and exp-log notation; and two function which calculate their derivatives.

2. Lang's MPH and HLP Models

J. B. Lang (2004) introduced a broad new class of contingency table models. This class is called Multinomial-Poisson Homogeneous (MPH) models, which can be characterized

by a sampling plan ($\mathbf{Z}$, $\mathbf{Z}_F$, $\mathbf{n}_F$) and a system of homogeneous, sufficiently smooth, constraints $\mathbf{h}(\mathbf{m}) = \mathbf{0}$, where $\mathbf{m}$ is the vector of expected table counts. Let $\mathbf{Z}$ be a (c x s) population matrix and let $\mathbf{Z}_F = \mathbf{Z}\mathbf{Q}_F$ be a sampling constraint matrix, the sampling plan is characterized by the triple where $\mathbf{Z}$ determines the strata from which samples are drawn, $\mathbf{Z}_F$ indicates which samples have a priori fixed sample size, the non fixed sample sizes have Poisson distributions and $\mathbf{n}_F$ gives the collection of fixed sample sizes, and the collection of unknown expected sample sizes is denoted $\boldsymbol{\delta}$ and the entire collection of expected sample sizes by $\boldsymbol{\gamma}$. We consider the joint probability function of q variables. The vector of the joint probabilities will be denoted by the vector $\boldsymbol{\pi}$ and its probabilities are called pre-sample probabilities. The data model probabilities are denoted by $\mathbf{p}$: p_{is} is a conditional probability, given sample from stratum s, the chance of a type i outcome. The vector of conditional probabilities satisfies this property:

$$\mathbf{p} = \mathbf{D}^{-1}\left(\mathbf{Z}\mathbf{Z}'\boldsymbol{\pi}\right)\boldsymbol{\pi} > 0 \tag{2.1}$$

The likelihood function is assumed to be the product of f multinomial probability functions depending on the elements of $\mathbf{p}$ and s - f Poisson probability functions depending on the elements of $\boldsymbol{\delta}$**.** It follows that $\mathbf{y}$, the vector of observed counts, is said to have a product multinomial-Poisson distribution with respect to the sampling plan.

There is a one-to-one correspondence between the ($\boldsymbol{\gamma}$, $\mathbf{p}$) parameters and the expected counts or mean parameter: $\mathbf{m} = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma}.)\mathbf{p}$ and $\mathbf{p} = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}'\mathbf{m})\mathbf{m}$.

The probability density function of the *Multinomial-Poisson* (MP) random vector $\mathbf{y}$, can be parameterized in terms of ($\boldsymbol{\gamma}$, $\mathbf{p}$): MP_Z^* ($\boldsymbol{\gamma}$, $\mathbf{p}$ | $\mathbf{Z}_F$, $\mathbf{n}_F$); or in terms of the expected count vector: MP_Z ($\mathbf{m}$ | $\mathbf{Z}_F$, $\mathbf{n}_F$). The ($\boldsymbol{\gamma}$, $\mathbf{p}$) parameterization is useful for the study of asymptotic behavior of model estimators, while the $\mathbf{m}$ parameterization is convenient for model fitting and specification.

A function $\mathbf{h}$: $\Omega \rightarrow \mathrm{R}^u$ is $\mathbf{Z}$-homogeneous of order $o(j)$, j = 1, …, u, if and only if for every component h_j of the function $\mathbf{h}$ it is:

$$h_j\left(\mathbf{D}\left(\mathbf{Z}\boldsymbol{\gamma}\right)\mathbf{p}\right) = \gamma_{v(j)}^{o(j)} h_j\left(\mathbf{p}\right) \tag{2.2}$$

where $\gamma_{v(j)}$ is a component of the vector $\boldsymbol{\gamma}$, $v_{(j)} \in \{1, 2, ..., s\}$.

Sufficient, but not necessary, condition for $\mathbf{Z}$-homogeneity is that if $\mathbf{h}$ is only a function of the expected counts $\mathbf{m}$ through the cell probabilities $\mathbf{p}$, then $\mathbf{h}$ is $\mathbf{Z}$-homogeneous. In words, if $\mathbf{h}$ is only a function of the expected counts $\mathbf{m}$ through the cell probabilities then $\mathbf{h}$ is $\mathbf{Z}$-homogeneous. Necessary, but not sufficient, condition for $\mathbf{Z}$-homogeneity: if $\mathbf{h}$ is $\mathbf{Z}$-homogeneous then $\mathbf{h}(\mathbf{m}) = \mathbf{0}$ if and only if $\mathbf{h}(\pi(\mathbf{m})) = \mathbf{0}$. In words, if $\mathbf{h}$ is $\mathbf{Z}$-homogeneous then constraining $\mathbf{m}$ via $\mathbf{h}(\mathbf{m}) = \mathbf{0}$ is equivalent to constraining $\mathbf{p}$ via $\mathbf{h}(\mathbf{p}) = \mathbf{0}$. An important special case of $\mathbf{Z}$-homogeneous function is the following zero order function: $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{p}) = \mathbf{h}(\mathbf{p})$. Lang's propositions describe the main consequences of these definitions that lead to simplification in model fitting and in derivations of the model equivalence results.

A special case of these models are the so called HLP models for contingency tables, which constraint the expected table counts through $\mathbf{L}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{L}$ is a sufficiently smooth link function that is allowed to be many-to-one and nonlinear, $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ is a vector of unknown regression parameters, together $\mathbf{X}\boldsymbol{\beta}$ is referred to as the linear predictor. The HLP class is very broad and includes models that are not of the

univariate and multivariate generalized linear model form. Most linear predictor contingency table models used in practice are member of this class. The constraint must satisfy the following conditions:

$$\begin{aligned}&\mathbf{L}(\mathbf{m}) = a(\boldsymbol{\gamma}) + \mathbf{L}(\mathbf{p}),\ \ a(\gamma_1) - a(\gamma_2) = a(\gamma_1/\gamma_2) - a(\mathbf{1});\\ &\mathbf{h}(\mathbf{m}) = \mathbf{U}'\mathbf{L}(\mathbf{m}) \text{ is sufficiently smooth and } \mathbf{Z}\text{-homogeneous.}\end{aligned} \tag{2.3}$$

Lang(1996) considers maximum likelihood methods for a broad class of models useful for describing multivariate categorical response data. These models, which are referred to as generalized log-linear models GLLM's., can be specified in terms of the vector of cell probabilities $\boldsymbol{\pi}$ as $\mathbf{C}\log \mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}$. Standard log- and logit-linear models cumulative and adjacent-category logit models for marginal distributions and global cross-ratios models are all special cases of these generalized log-linear models.

3. The Goodman and Kruskal tau-*b* and the Gini's index

Tau-*b* is a widely used measure of association for categorical data, introduced by Goodman and Kruskal has been proposed with a different interpretation by Light and Margolin (1971, 1974). Although the method used to arrive at R^2 differs from the method used for τ_b and the interpretation is also different, Margolin's R^2 is algebraically equivalent to the Goodman and Kruskal's τ_b. Tau-*b* is related to a well know categorical data measure of variation: the Gini's index.
In order to use the fitting and testing methods, the derivative matrix of these measures must be calculated then we introduce the Goodman and Kruskal τ_b and the Gini's index in matrix notation. For this, we follow the approach of Kritzer (1977), who gave matrix formulas for several measures of association and the generalization introduced by Bergsma (1997). The method of matrix notation that we propose has been referred to as the "exp-log" notation.
Following Lang it is useful to express the constraints using expected counts **m**. While to easily find the **h**(**m**) derivative, for the ML estimation, we need to write them in exp-log notation.

$$\tau_{X_1/X_2} = \exp\left\{ \underset{(1\times 2)}{GK}\mathbf{C}_1' \cdot \log\left[\underset{(2\times(IJ+I+1))}{GK}\mathbf{A}_1' \cdot \left(\exp\left\{ \underset{((IJ+I+1)\times(IJ+I+J+1))}{GK}\mathbf{C}' \cdot \log\left[\underset{((IJ+I+J+1)\times IJ)}{GK}\mathbf{A}' \cdot \mathbf{m} \right] \right\} \right) \right] \right\} \tag{3.1}$$

The Goodman and Kruskal's τ, using the recursive definition and the expected counts is still an $\mathbf{h}_2$-function.

$$V(X_1) = \underset{1\times(I+1)}{\mathbf{e}}' \cdot \exp\left\{ \underset{((I+1)\times(I+1))}{G}\mathbf{C}' \cdot \log\left[\underset{((I+1)\times IJ)}{G}\mathbf{A}' \cdot \mathbf{m} \right] \right\} \tag{3.2}$$

The Gini's index, using the recursive definition and the expected counts is still an $\mathbf{h}_1$function. Both the link functions considered are zero order **Z**-homogeneous: formulas are identical using probabilities or expected counts and this is a special case that simplify inference too.

For the applications presented we write some R functions, to define the constraints in the HLP models and to find their analytic derivatives. We use hmmm R's package.
We present two motivating example, in the first dataset consider occupants involved in car accident injuries classified according to injury gravity, restraint usage and year. We model five tau-*b* directly as a function of time. We test the hypothesis of no changing association in the five years and also the Gini's index as link function. In the second motivating example we use data were obtained from a multicenter randomized clinical trial involving suitably eligible patients who were treated in four participating hospitals. We test the hypothesis of no changing association in the different hospitals, because model obtained could be compared to an additive model: Grizzle, Starmer and Koch (1969), using additive model, found that there are no significant hospital effects too. Forthofer and Koch (1973) used them in an analysis of rank correlation, where they use as compounded functions the tau-*b.* Interpreting the same model as a measure of the partial association between the severity of the dumping syndrome and the extent of the operation they had our same results too. An interesting case presented is that of an HLP model specified by simultaneous constraints.

4. Conclusion

Using this approach we obtain also approximate 95% confidence intervals and test statistics, G^2, X^2, W^2, for assessing the overall goodness of fit, differently than with other approach, we obtain the estimated expected counts under the assumed hypothesis.
As Lang we assumed that the ML estimates exist and uniquely solve the restricted likelihood equations. Unfortunately, to prove uniqueness or existence results, we use results for "regular" marginal models described by Bergsma (1997). Outside this and other special settings, however, the ML estimates may not exist, the likelihood equations may have several solution, and, or the ML estimate could be nonunique . The existence und uniqueness issue for the broad class of HLP models is an unsolved problem, and remains a topic for future research.

References

Bergsma W. P. (1997) *Marginal Models for Categorical Data*. Tilburg University Press, Tilburg.

Cazzaro M., Colombi R. (2007) Hierarchical Multinomial Marginal Models and the R-package hmmm: a brief introduction, www.unibg.it/pers/colombi.

Gini C. W. (1912) Variabilità e mutabilità. *Studi Economico-Giuridici dell'Università di Cagliari*, 3, 3-159.

Goodman L. A., Kruskal W. H. (1979) *Measures of Association for Cross Classifications*. Springer-Verlag, New York.

Kritzer H. M. (1977) Analyzing measures of association derived from contingency tables. *Sociological Methods & Research*, Sage Publications, 5(4), 387-418.

Lang J. B. (2004) Multinomial-Poisson homogeneous models for contingency tables. *The Annals of Statistics*, 32(1), 340-383.

Lang J. B. (2005) Homogeneous linear predictor models for contingency tables. *Journal of the American Statistical Association*, 100(469-Theory and Methods), 121-134.