# Log-mean linear regression models for assessing the effect of HIV-infection on multimorbidity in a case-control study

## *Modelli di regressione log-mean linear per valutare l'effetto del virus HIV sulla multimorbidità in uno studio caso-controllo*

Monia Lupparelli and Alberto Roverato

**Abstract** We propose a multivariate regression approach for binary variables based on the log-mean linear link function for the response variables. This approach has shown to provide useful insights for assessing the effect of HIV-infection on multimorbidity, defined as the occurrence of co-existing noninfectious diseases for a sample of patients deriving from a case-control study. The coefficients of these regression models are log-linear combinations of relative risks and we show that submodels identified by zero regression coefficients encode relevant hypotheses for the considered application.

**Abstract** Proponiamo un metodo di regressione multivariata per variabili binarie basato sulla funzione link log-mean linear per le variabili risposta. Questo approccio è molto utile per valutare l'effetto del virus HIV sulla multimorbidità, intesa come il verificarsi congiunto di più patologie non infettive in un campione di individui derivanti da uno studio caso-controllo. I coefficienti di regressione sono combinazioni log-lineari di rischi relativi e sottomodelli definiti da zero parametri definiscono ipotesi di interesse in questa applicazione.

**Key words:** link function, disease pattern, relative risk, response interaction

## 1 Introduction

The co-occurrence of two or more chronic medical conditions in one person is generally defined as *multimorbidity*. It is well-known that multimorbidity correlates with age and, furthermore, that patients affected by HIV show an increased prevalence of noninfectious comorbidities, compared with the general population. The

Monia Lupparelli
Department of Statistical Sciences, Via Belle Arti, 41, Bologna, e-mail: monia.lupparelli@unibo.it

Alberto Roverato
Department of Statistical Sciences, Via Belle Arti, 41, Bologna, e-mail: alberto.roverato@unibo.it

effect of HIV-infection on noninfectious diseases was studied by [3] in a cross-sectional retrospective case-control study with 25% of cases and 75% of controls. Patients affected by HIV and older than 18 were consecutively enrolled at the Metabolic Clinic of Modena University (IT) from 2002 to 2009. The study was conducted including control subjects matched according to some criteria (e.g. age and gender) and selected from the CINECA ARNO Observational Database. Let $Y_b$ (*Bone Fracture*), $Y_d$ (*Diabete Mellitus*), $Y_c$ (*Cardiovascular Disease*) and $Y_r$ (*Renal Disease*) be four binary random variables representing noninfectious diseases taking value 1 for the occurrence of each disease and 0 otherwise. [3] analyzed the effect of HIV-infection by means of logistic regressions on univariate responses representing single diseases. Nevertheless, multimorbidity is characterized by the complex interaction of co-existing diseases, then the role of HIV can be better explored through a multivariate regression approach rather than univariate ones. Our intent is investigating the effect of HIV on disease patterns represented by means of interactions of responses. Therefore, we are mainly interested in a multivariate regression approach for modelling the joint dependence of response variables on a set of covariates.

The statistical properties of a broad family of linear predictor models for multivariate categorical responses were investigated by [4] who also implemented an algorithm for maximum likelihood estimation. This family of models has systematic components: a link function of the observed cell counts deriving from a Multinomial-Poisson sampling scheme, a regression design generally defined through a design matrix and a set of regression coefficients. From the model interpretability perspective, the choice of a suitable link function is a crucial aspect because it defines (i) the response interaction and, therefore, (ii) the regression coefficient interpretation. In particular, we discuss and illustrate the main features of two regression approaches based on two distinct link functions which both belong to the class of models defined in [4].

For our analysis, we preliminary consider the class of regression models defined by the *log-mean* (LM) link function also used by [1] to define chain graph models. In LM regressions (i) the response interactions are the log-probability of the occurrence of disease patterns and (ii) the regression coefficients are the log-relative risk of these occurrences given the exposure to the HIV risk factor. This approach, even if appealing for the parameter interpretability, might be limited for model selection because testing zero regression coefficients does not correspond to hypotheses of interest in this context. Then, we propose a regression approach based on the *log-mean linear* (LML) link function inspired by the parameterization recently developed by [6]. Setting the regression coefficients to zero in LML regressions defines models of interest which have a natural interpretation in terms of additive log-relative risks.

## 2 The LM regression model

Let $Y_V = (Y_v)_{v \in V}$ be a binary random vector of response variables taking values $y_V \in \mathscr{I}_V$, with $\mathscr{I}_V = \{0,1\}^{|V|}$ and let $H$ be a binary covariate taking values $h \in \mathscr{I}$,

with $\mathscr{I} = \{0,1\}$. Let $Z_V$ denote a square matrix of size $2^{|V|}$ with entries indexed by the subsets of $V \times V$ such that $Z_{D,F} = 1(D \subseteq F)$; its inverse $M_V = Z_V^{-1}$ is called the Möbius matrix, see [6] for technical details. For $|V| = 1$, we adopt the short notation $Z$ and $M$ without the subscript for the corresponding $(2 \times 2)$ square matrices.

Each random variable $Y_V|\{H = h\}$ follows a multivariate Bernoulli distribution and let $\pi = \{\pi_D(h)\}_{D \subseteq V, h \in \mathscr{I}}$ be the matrix where each column vector $\pi(h)$ of size $2^{|V|}$ is the strictly positive probability parameter of $Y_V|\{H = h\}$ with generic element $\pi_D(h) = P(Y_D = 1_D, Y_{\bar{D}} = 0_{\bar{D}}|H = h)$. The mean parameter of the same distribution is given by the column vector $\mu(h)$ of the matrix $\mu = Z_V \pi$ with generic element $\mu_D(h) = P(Y_D = 1_D|H = h)$.

The saturated LM regression model for the conditional distribution of $Y_V|H$ is characterized by the expansion

$$\log(\mu_D(h)) = \sum_{h' \in \mathscr{I}: h' \leq h} \beta_D^{\langle \mu \rangle}(h'), \qquad h \in \mathscr{I}, \tag{1}$$

for every $D \subseteq V$. Let $\beta^{\langle \mu \rangle} = \log(\mu)M$ be the matrix of regression coefficients where, for each subset $D$ of $V$, the generic elements are $\beta_D^{\langle \mu \rangle}(0) = \log P(Y_D = 1_D|H = 0)$ and $\beta_D^{\langle \mu \rangle}(1)$ which is the log-relative risk

$$\log RR_D(H) = \log \frac{P(Y_D = 1_D|H = 1)}{P(Y_D = 1_D|H = 0)}. \tag{2}$$

This model is based on a regression design also used by [2] to define multivariate logistic regression models which still belong to the regression family of [4]; see [5] for an application to regression chain graphs.

For the multimorbidity analysis which involves four diseases $Y_V = \{Y_b, Y_d, Y_c, Y_r\}$ as response variables and the covariate $H$ (taking value 1 when the HIV-infection is present and 0 otherwise), the LM regression approach is appealing because the model parameters represent measures of interest: (i) the response interactions are the log-probabilities of multimorbidity events, i.e. the occurrence of different disease patterns, and (ii) the coefficients $\beta^{\langle \mu \rangle}$ measure the effect of HIV on each pattern in terms of log-relative risk.

Nevertheless, this approach might be limited for model selection because, for any disease pattern $Y_D$ with $D \subseteq V$, testing both the null model by means of $\beta_D^{\langle \mu \rangle}(0) = \beta_D^{\langle \mu \rangle}(1) = 0$ and the fixed effect model by means of $\beta_D^{\langle \mu \rangle}(1) = 0$ is not of interest in this contest. The former would imply a zero value for the log-mean probability $\log(\mu_D(h))$ for any $h \in \mathscr{I}$ and, therefore, a not strictly positive probability parameter $\pi(h)$ which is not feasible for the distribution of $Y_V|\{H = h\}$. The latter would imply the HIV-infection has no effect on the occurrence of disease patterns, a hypothesis which is extremely rare because we *a priori* know the HIV has a role on these multimorbidity events and our main interest is in assessing this effect. Hence, we propose a LML link function for our regression modelling based on the parameterization developed by [6]. In LML regressions both the null and the fixed effect models define hypotheses of interest which are useful for model selection.

## 3 LML regression models for multimorbidity analysis

The saturated LML regression model for the conditional distribution of $Y_V|H$ is characterized by the expansion

$$\gamma_D(h) = \sum_{h' \in \mathscr{I}: h' \leq h} \beta_D^{\langle\gamma\rangle}(h'), \qquad h \in \mathscr{I}, \tag{3}$$

for every $D \subseteq V$. The link function is $\gamma_D(h) = \sum_{D' \subseteq D} (-1)^{|D \setminus D'|} \log(\mu_{D'}(h))$ and $\beta^{\langle\gamma\rangle} = \gamma M$ is the matrix of regression coefficients. Exploiting the linear mapping between the LML and the LM parameterization, it can be easily verified that there is also a linear mapping between the coefficients of the LML and the LM regression; see [6] for technical details about the LML parameterization and the mapping with the LM one. Then,

$$\beta^{\langle\mu\rangle} = Z_V' \beta^{\langle\gamma\rangle}, \tag{4}$$

such that, for every $D \subseteq V$ and $h \in \mathscr{I}$,

$$\beta_D^{\langle\mu\rangle}(h) = \sum_{D' \subseteq D} \beta_{D'}^{\langle\gamma\rangle}(h), \qquad \beta_D^{\langle\gamma\rangle}(h) = \sum_{D' \subseteq D} (-1)^{|D \setminus D'|} \beta_{D'}^{\langle\mu\rangle}(h). \tag{5}$$

We remark that the linear mapping between these regression coefficients is an important and a peculiar feature of the LML regression approach which derives from the fact that the inverse mapping between the LML parameterization and the probability parameter can be explicitly written in analytic form; see [6]. This property is not so common in parameterizations for categorical variables, with the exception of some relevant cases as the well-known log-linear parameterization.

The interpretation of the LML regression model is less intuitive than the LM one because (i) the response interactions are log-linear combinations of mean probabilities and (ii) the regression coefficients are log-linear combinations of relative risks. Nevertheless, model selection within the LML approach can be easily implemented testing for each $Y_D|H$, with $D \subseteq V$, both the null and the fixed effect model which have a straightforward and useful interpretation in the multimorbidity analysis.

A *null LML regression model*, defined by $\beta_D^{\langle\gamma\rangle}(0) = \beta_D^{\langle\gamma\rangle}(1) = 0$ for any $D \subseteq V$, implies $\gamma_D(0) = \gamma_D(1) = 0$, and, in the special case of $|D| = 2$, a conditional independence model. For instance, let $D = \{b, d\}$, the null model implies $Y_b \perp\!\!\!\perp Y_d|H$; see [6] for a proof.

A *fixed effect LML regression model*, defined by $\beta_D^{\langle\gamma\rangle}(1) = 0$ for any $D \subseteq V$, implies an additive log-relative risk for the outcome $Y_D$ because, from (5) follows that $\beta_D^{\langle\mu\rangle}(1) = \sum_{D' \subset D} \beta_{D'}^{\langle\gamma\rangle}(1)$. For instance, let $D = \{b, d\}$, the fixed effect model implies

$$\log RR_{\{b,d\}}(H) = \log RR_b(H) + \log RR_d(H). \tag{6}$$

Assumption in (6) is weaker than conditional independence, because $Y_b$ and $Y_d$ are not independent either in cases or in controls, however, the interaction between $Y_b$ and $Y_d$ is the same both in cases and in controls.

**Table 1** The fitted LML regression model, $\chi^2_{12} = 3.448$ (0.991). Let $\cdot$ denotes the zero coefficients.

| Model | $\hat{\beta}_D^{\langle\gamma\rangle}(0)$ | s.e | p-value | $\hat{\beta}_D^{\langle\gamma\rangle}(1)$ | s.e | p-value | $\hat{\beta}_D^{\langle\mu\rangle}(0)$ | s.e | $\hat{\beta}_D^{\langle\mu\rangle}(1)$ | s.e | $w_D$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y_b \sim H$ | -4.534 | 0.088 | 0.000 | 2.579 | 0.100 | 0.000 | -4.534 | 0.088 | 2.579 | 0.100 | 0.3228 |
| $Y_d \sim H$ | -3.253 | 0.053 | 0.000 | 1.058 | 0.075 | 0.000 | -3.253 | 0.053 | 1.058 | 0.075 | 0.4223 |
| $Y_c \sim H$ | -4.474 | 0.093 | 0.000 | 1.056 | 0.138 | 0.000 | -4.474 | 0.093 | 1.056 | 0.138 | 0.1241 |
| $Y_r \sim H$ | -6.131 | 0.089 | 0.000 | 3.383 | 0.114 | 0.000 | -6.131 | 0.089 | 3.383 | 0.114 | 0.1308 |
| $Y_{bd} \sim H$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | -7.787 | 0.101 | 3.637 | 0.122 | 0.2536 |
| $Y_{bc} \sim H$ | 0.443 | 0.181 | 0.014 | $\cdot$ | $\cdot$ | $\cdot$ | -8.565 | 0.227 | 3.635 | 0.178 | 0.1153 |
| $Y_{br} \sim H$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | -10.665 | 0.011 | 5.962 | 0.086 | 0.1353 |
| $Y_{dc} \sim H$ | 1.777 | 0.124 | 0.000 | -1.232 | 0.235 | 0.000 | -5.950 | 0.182 | 0.882 | 0.291 | 0.2102 |
| $Y_{dr} \sim H$ | 0.840 | 0.102 | 0.000 | $\cdot$ | $\cdot$ | $\cdot$ | -8.544 | 0.155 | 4.440 | 0.142 | 0.2550 |
| $Y_{cr} \sim H$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | -10.605 | 0.118 | 4.439 | 0.172 | 0.0306 |
| $Y_{bdc} \sim H$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | -10.041 | 0.261 | 3.461 | 0.299 | 0.2229 |
| $Y_{bdr} \sim H$ | 1.720 | 0.112 | 0.000 | -1.382 | 0.238 | 0.011 | -11.358 | 0.011 | 5.638 | 0.251 | 0.4991 |
| $Y_{bcr} \sim H$ | 3.337 | 0.202 | 0.000 | -2.631 | 0.431 | 0.000 | -11.358 | 0.006 | 4.387 | 0.439 | 0.1391 |
| $Y_{dcr} \sim H$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | -11.241 | 0.222 | 4.264 | 0.323 | 0.1389 |
| $Y_{bdcr} \sim H$ | -1.777 | 0.124 | 0.000 | 1.284 | 0.530 | 0.015 | -12.051 | 0.006 | 4.115 | 0.729 | 1.0000 |
| $H$ | -1.387 | 0.016 | 0.000 | | | | -1.387 | 0.016 | | | |

We fitted the data using a LML regression approach and maximum likelihood estimations have been derived by means of the `mph package` implemented in `R` by J. Lang and available on his web-site upon request. We selected a model with a deviance $\chi^2_{(12)} = 3.448$ whose nonzero $\beta^{\langle\gamma\rangle}$ coefficients are collected in the left side block of Table 1. The model has been selected using a backward stepwise procedure which, starting from the saturated model, selects sub-models with good fitting (p-value $\geq 0.05$ of the deviance) setting zero non-significant coefficients until all the remaining ones are significant (p-value $< 0.05$).

The selected model implies additive log-relative risks for the disease patterns $\{Y_b, Y_d\}$, $\{Y_b, Y_r\}$, $\{Y_c, Y_r\}$, $\{Y_b, Y_c\}$, $\{Y_d, Y_r\}$, $\{Y_b, Y_d, Y_c\}$ and $\{Y_d, Y_c, Y_r\}$, and the pairwise conditional independence for the first three patterns. The middle side block of Table 1 includes the corresponding log-relative risks $\beta^{\langle\mu\rangle}$ (LM regression coefficients) for the fitted model computed on the basis of (4). All of them are nonzero and the log-relative risks are positive, as expected. The lower relative risk is for the pattern $\{Y_d, Y_c\}$, whereas the highest ones are for the patterns $\{Y_b, Y_r\}$ and $\{Y_b, Y_d, Y_r\}$.

In order to better assess the role of HIV-infection on multimorbidity, we derive a set of indexes $\rho_k$, with $k = 1, \ldots, 4$ representing the disease pattern size, which measure the HIV effect for the co-occurrence in patients of one, two, three or four diseases, respectively. Every $\rho_k$ is the average log-relative risk computed among disease patterns of size $k$ weighted with the probability of the occurrence of each pattern. These weights, denoted by $w_D$, with $D \subseteq V$, are the estimated probabilities of the occurrence of each disease pattern normalized within the group of patterns of the same size. They are shown in the right side block of Table 1 and, in details, for every pattern $D \subseteq V$, $w_D$ is the estimate of $P(Y_D = 1_D)$ normalized with respect to the sum of all estimated $P(Y_{D'} = 1_{D'})$, for each $D' \subseteq V$ such that $|D'| = |D|$. Then, for every $D \subseteq V$, $\sum_{D' \subseteq V : |D'| = |D|} w_{D'} = 1$.

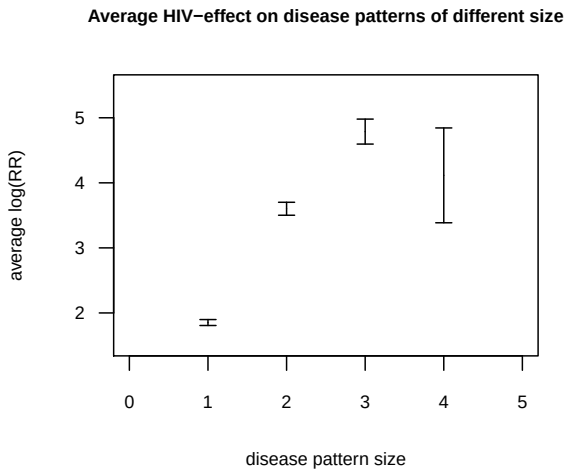


**Fig. 1** Confidence intervals for the indexes $\rho_k$, with $k = 1, \ldots, 4$.

The value of these indexes with their standard errors included in brackets are $\rho_1 = 1.853(0.046)$, $\rho_2 = 3.602(0.100)$, $\rho_3 = 4.788(0.192)$ and $\rho_4 = 4.115(0.729)$. Confidence intervals for these indexes are displayed in Figure 1. The plot shows the confidence intervals for $k = 1, 2$ are more accurate than for $k = 3$ and, in particular, for $k = 4$, because in the observed sample the number of patients affected by one or two diseases is larger than the number of patients affected by three or four diseases. However, we can conclude the average HIV-effect on multimorbidity events of different size $k$ is never close to zero and the risk of the occurrence of co-existing diseases for patients affected by HIV-infection in average increases with the disease pattern size.

## References

1. Drton, M.: Discrete chain graph models. Bernoulli **15**, 736-753 (2009)
2. Glonek, G.J.N., McCullagh, P.: Multivariate logistic models. Journal of the Royal Statistical Society, Series B (Methodological) **57**, 533-546 (1995)
3. Guaraldi, G., Orlando, G., Zona, S., Menozzi, M., Carli, F., Garlassi, E., Berti, A., Rossi, E., Roverato, A., Palella, F.: Premature age-related comorbidities among HIV-infected persons compared with the general population. Clinical Infectious Diseases **53**, 1120-1126 (2011)
4. Lang, J.B,: Homogenous linear predictor model for contincency tables. Journal of the American Statistical Association **100**, 121–134 (2005)
5. Marchetti, G.M., Lupparelli, M.: Chain graph models of multivariate regression type for categorical data. Bernoulli. **17**, 827-844 (2011)
6. Roverato, A., Lupparelli, M., La Rocca, L.: Log-mean linear models for binary data. Biometrika **100**, 485-494 (2013)