UNIVERSITÀ DEGLI STUDI FIRENZE

FLORE
Repository istituzionale dell'Università degli Studi di Firenze

# Log-mean linear regression models for binary responses with an application to multimorbidity.

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

*Original Citation:*

Log-mean linear regression models for binary responses with an application to multimorbidity / Lupparelli, M., Roverato, A.. - In: JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS. - ISSN 0035-9254. - STAMPA. - 66:(2017), pp. 227-252.

*Availability:*

This version is available at: 2158/1138477 since: 2018-10-30T10:02:02Z

*Terms of use:*

Open Access

La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf)

*Publisher copyright claim:*

(Article begins on next page)

09 August 2024

arXiv:1410.0580v3 [stat.ME] 16 May 2016

# Log-mean linear regression models for binary responses with an application to multimorbidity

Monia Lupparelli* and Alberto Roverato†

March 30, 2016

**Abstract**

In regression models for categorical data a linear model is typically related to the response variables via a transformation of probabilities called the link function. We introduce an approach based on two link functions for binary data named *log-mean* (LM) and *log-mean linear* (LML), respectively. The choice of the link function plays a key role for the interpretation of the model, and our approach is especially appealing in terms of interpretation of the effects of covariates on the association of responses. Similarly to Poisson regression, the LM and LML regression coefficients of single outcomes are log-relative risks, and we show that the relative risk interpretation is maintained also in the regressions of the association of responses. Furthermore, certain collections of zero LML regression coefficients imply that the relative risks for joint responses factorize with respect to the corresponding relative risks for marginal responses. This work is motivated by the analysis of a dataset obtained from a case-control study aimed to investigate the effect of HIV-infection on *multimorbidity*, that is simultaneous presence of two or more noninfectious commorbidities in one patient.

*Keywords*: Categorical data; Link function; Multimorbidity pattern; Relative risk; Response association.

*Department of Statistical Sciences, University of Bologna, IT (monia.lupparelli@unibo.it)
†Department of Statistical Sciences, University of Bologna, IT (alberto.roverato@unibo.it)

# 1 Introduction

In many research fields it is required to model the dependence of a collection of response variables on one or more explanatory variables. McCullagh and Nelder (1989, Sec. 6.5) specified that in this context there are typically three lines of inquiry: (i) the dependence structure of each response marginally on covariates, (ii) a model for the joint distribution of all responses and (iii) the joint dependence of response variables on covariates. When in a regression model responses are categorical, a linear model is typically related to the response variables via a transformation of probabilities called the link function, and the choice of the link function plays a key role for the interpretation of the model along the lines (i) to (iii). We refer to Tutz (2011) and Agresti (2013) for a full account of regression models for categorical data; see also Ekholm et al. (2000, Section 5) for a review of some link functions commonly used in the binary case.

This work is motivated by a research aimed to investigate the effect of HIV-infection on *multimorbiditiy* which is defined as the co-occurrence of two or more chronic medical conditions in one person. It is well known that multimorbidity is associated with age and, furthermore, that HIV-infected patients experience an increased prevalence of noninfectious comorbidities, compared with the general population. Guaraldi et al. (2011) considered a dataset obtained from a cross-sectional retrospective case-control study, and investigated the effect of HIV-infection on the prevalence of a set of noninfectious chronic medical conditions by applying an univariate regression to each response. However, multimorbidity is characterised by complex interactions of co-existing diseases and to gain relevant insight it is necessary to use a multivariate approach aimed to investigate the effect of HIV on the way different chronic conditions associate. The main scientific objective of this study is thus the line of enquiry (iii). However, to the best of our knowledge, this line has never been explicitly addressed in the literature, and this paper is fully devoted to this issue.

The application we consider naturally requires a *marginal modelling* approach because the main interest is for the effect of HIV on the marginal association of subsets of comorbidities; see Tutz (2011, Chapter 13) and Agresti (2013, Chapter 12). For this reason, we focus on the case where the link function satisfies *upward compatibility*, that is every association term among responses can be computed in the relative marginal distribution. In this way, the parameterization of the response variables will include terms that can be regarded as *single outcomes*, computed marginally on univariate responses, and terms that can be regarded as *association outcomes*, hereafter referred

to as *response associations*, which are computed marginally on subsets of responses. Regression models typically include coefficients encoding the effect of the covariates, as well as of interactions of covariates, on response associations and difficulties involve both the interpretation of the response associations and the interpretation of the relevant regression coefficients. More seriously, the effect of a covariate on a response association might be *removable* in the sense that it disappears when a different link function is used. It is therefore crucially relevant to be able to define models with interpretable regressions coefficients; see Berrington de González and Cox (2007) for a review on statistical interactions, with emphasis on interpretation.

In marginal modelling a central role is played by the *multivariate logistic* regression model because it maintains a marginal logistic regression interpretation for the single outcomes (McCullagh and Nelder, 1989; Glonek and McCullagh, 1995). Nevertheless, this family of regression models does not provide a satisfying answer to the multimorbidity application. This is due to the fact that, although the regression coefficients for the single outcomes can be interpreted in terms of odds ratios, this feature does not translate to the higher order regressions where both the response association and the relevant regression coefficients are high-level log-linear parameters which are difficult to interpret.

We consider two different parameterizations, the *log-mean* (LM) (Drton and Richardson, 2008; Drton, 2009) and the *log-mean linear* (LML) parameterization (Roverato et al., 2013; Roverato, 2015), and investigate the use of these parameterizations as link functions. In this way, we introduce an approach where, similarly to Poisson regression, regression coefficients can be interpreted in terms of relative risks. Furthermore, and more interestingly, the relative risk interpretation can be extended from the regressions of the single outcomes to the regressions of the response associations, thereby providing interpretable coefficients. The LM and the LML links can be used to specify the same classes of submodels but the LML link has the advantage that relevant submodels can be specified by setting regression coefficients to zero. Specifically, we show that certain collections of zero LML regression coefficients imply that the relative risks for joint responses factorize with respect to the corresponding relative risks for marginal responses.

The paper is organized as follows. In Section 2 we describe the motivating problem concerning the analysis of the multimorbidity data. Section 3 gives the background concerning the theory of regression for multivariate binary responses, as required for this paper. Section 4 introduces the LM and the LML regression models and describes the relevant properties of these models. The analysis of multimorbidity data is carried

out in Section 5 and, finally, Section 6 contains a discussion.

# 2 Motivating problem: multimorbidity in HIV-positive patients

Antiretroviral therapy (ART) for human immunodeficiency virus (HIV) infection has been a great medical success story. Nowadays, in countries with good access to treatment, clinical AIDS is no longer the inevitable outcome of HIV infection and this disease, previously associated with extremely high mortality rates, is now generally thought of as a chronic condition (Mocroft et al., 2003; May and Ingle, 2011). Despite a marked increase in life expectancy, mortality rates among HIV-infected persons remain higher than those seen in the general population. Some of the excess mortality observed among HIV-infected persons can be directly attributed to illnesses that occur as a consequence of immunodeficiency, however, more than half of the deaths observed in recent years among ART-experienced HIV infected patients are attributable to noninfectious comorbidities (NICMs) (Phillips et al., 2008; Guaraldi et al., 2011).

*Multimorbidity* is defined as the co-occurrence of two or more chronic medical conditions in one person that is, for HIV positive patients, as the simultaneous presence of two or more NICMs. Multimorbidity, which is associated with age, is perhaps the most common "disease pattern" found among the elderly and, for this reason, it is turning into a major medical issue for both individuals and health care providers (Marengoni et al., 2011). It is well known that HIV-infected patients experience an increased prevalence of NICMs, compared with the general population, and it has been hypothesized that such increased prevalence is the result of premature aging of HIV-infected patients (Deeks and Phillips, 2009; Shiels et al., 2010; Guaraldi et al., 2011). Multimorbidity is characterised by the co-occurrence of NICMs and therefore investigating the effect of HIV-infection on multimorbidity requires to investigate the effect of HIV on the way different chronic conditions associate.

The dataset we analyse here comes from a study of Guaraldi et al. (2011) who investigated the effect of HIV-infection on the prevalence of a set of noninfectious chronic medical conditions. Data were obtained from a cross-sectional retrospective case-control study with sample size $n = 11\,416$ (2854 cases and 8562 controls). Cases were ART-experienced HIV-infected patients older than 18 years of age who were consecutively enrolled at the Metabolic Clinic of Modena University in Italy from 2002 to 2009. Control subjects were matched according to age, sex, race (all white),

and geographical area. The observed variables include both a set of binary (response) variables encoding the presence of NICMs of interest and a set of context and clinical covariates. See Guaraldi et al. (2011) for details and additional references.

# 3 Background and notation

## 3.1 Möbius inversion

In this subsection we introduce the notation used for matrices and recall a well known result named *Möbius inversion* that will be extensively used in the following.

For two finite sets $V$ and $U$, with $|V| = p$ and $|U| = q$, we write $\theta = \{\theta_D(E)\}_{D\subseteq V, E\subseteq U}$ to denote a $2^p \times 2^q$ real matrix with rows and columns indexed by the subsets of $V$ and $U$, respectively. Furthermore, we will write $\theta(E)$ to denote the column of $\theta$ indexed by $E \subseteq U$ and $\theta_D$ to denote the row of $\theta$ indexed by $D \subseteq V$. Note that the notation we use may be easier to read if one associates $D$ with $D$iseases and $E$ with $E$xposure.

**Example 3.1 (Matrix notation.)** *For the case $V = \{b, c, d\}$ and $U = \{h, a\}$ the matrix $\theta$ has eight rows indexed by the subsets $\emptyset$, $\{b\}$, $\{c\}$, $\{d\}$ $\{b, c\}$, $\{b, d\}$, $\{c, d\}$, $\{b, c, d\}$ and four columns indexed by the subsets $\emptyset$, $\{h\}$, $\{a\}$, $\{h, a\}$. The matrix $\theta$, with row and column indexes, is given below; note that we use the suppressed notation $\theta_{bc}(ha)$ to denote $\theta_{\{b,c\}}(\{h, a\})$, and similarly for the other quantities.*

$$
\theta = \begin{array}{c} \begin{array}{cccc} \emptyset & \{h\} & \{a\} & \{h,a\} \end{array} \\ \begin{pmatrix} \theta_\emptyset(\emptyset) & \theta_\emptyset(h) & \theta_\emptyset(a) & \theta_\emptyset(ha) \\ \theta_b(\emptyset) & \theta_b(h) & \theta_b(a) & \theta_b(ha) \\ \theta_c(\emptyset) & \theta_c(h) & \theta_c(a) & \theta_c(ha) \\ \theta_d(\emptyset) & \theta_d(h) & \theta_d(a) & \theta_d(ha) \\ \theta_{bc}(\emptyset) & \theta_{bc}(h) & \theta_{bc}(a) & \theta_{bc}(ha) \\ \theta_{bd}(\emptyset) & \theta_{bd}(h) & \theta_{bd}(a) & \theta_{bd}(ha) \\ \theta_{cd}(\emptyset) & \theta_{cd}(h) & \theta_{cd}(a) & \theta_{cd}(ha) \\ \theta_{bcd}(\emptyset) & \theta_{bcd}(h) & \theta_{bcd}(a) & \theta_{bcd}(ha) \end{pmatrix} \end{array} \begin{array}{c} \\ \emptyset \\ \{b\} \\ \{c\} \\ \{d\} \\ \{b,c\} \\ \{b,d\} \\ \{c,d\} \\ \{b,c,d\} \end{array}
$$

*Furthermore, the rows and columns of $\theta$ are denoted as follows.*

$$\theta = \begin{pmatrix} \theta_{\emptyset} \\ \theta_b \\ \theta_c \\ \theta_d \\ \theta_{bc} \\ \theta_{bd} \\ \theta_{cd} \\ \theta_{bcd} \end{pmatrix} = \begin{pmatrix} \theta(\emptyset) & \theta(h) & \theta(a) & \theta(ha) \end{pmatrix}$$

This matrix notation is not standard in the literature concerning categorical data but, in the case of regression models for binary data, it allows us to provide a compact representation of model parameters in matrix form and to compute alternative parameterizations, as well as regression coefficients, by direct application of Möbius inversion.

Let $\omega$ be another real matrix indexed by the subsets of $V$ and $U$. For a subset $D \subseteq V$ Möbius inversion states that

$$\theta_D(E) = \sum_{E' \subseteq E} \omega_D(E'), \;\; \forall E \subseteq U \quad \Leftrightarrow \quad \omega_D(E) = \sum_{E' \subseteq E} (-1)^{|E \setminus E'|} \theta_D(E'), \;\; \forall E \subseteq U; \tag{1}$$

see, among others, Lauritzen (1996, Appendix A). Let $Z_U$ and $M_U$ be two $(2^q \times 2^q)$ matrices with entries indexed by the subsets of $U \times U$ such that the entry of $Z$ indexed by the pair $E, H \subseteq U$ is equal to $1(E \subseteq H)$ and the corresponding entry of $M$ is equal to $(-1)^{|H \setminus E|} 1(E \subseteq H)$, where $1(\cdot)$ denotes the indicator function. Then, the equivalence (1) can be written in matrix form as $\theta_D = \omega_D Z_U \;\Leftrightarrow\; \omega_D = \theta_D M_U$ and Möbius inversion follows by noticing that $M_U = Z_U^{-1}$. Note that it is straightforward to extend this result to the matrices $\omega$ and $\theta$ as

$$\theta = \omega Z_U \quad \Leftrightarrow \quad \omega = \theta M_U \tag{2}$$

and, furthermore, that it makes sense to consider Möbius inversion also with respect to the columns of $\omega$ and $\theta$ so that $\theta = Z_V^{\top} \omega \;\Leftrightarrow\; \omega = M_V^{\top} \theta$.

## 3.2 Multivariate binary response models

Let $Y_V = (Y_v)_{v \in V}$ be a binary random vector of response variables with entries indexed by $V$ and $X_U = (X_u)_{u \in U}$ a vector of binary covariates with entries indexed by $U$. Without loss of generality, we assume that $Y_V$ and $X_U$ take value in $\{0,1\}^p$ and $\{0,1\}^q$, respectively. The values of covariates denote different observational or experimental conditions and we assume that, for every $x_U \in \{0,1\}^q$, the distribution of $Y_V|(X_U = x_U)$ is multivariate Bernoulli. Furthermore, we assume that the latter distributions are independent across conditions and that, when $X_U$ is regarded as a random vector, then also $(Y_V, X_U)$ follows a multivariate Bernoulli distribution. We can write the probability distributions of $Y_V|X_U$ by means of a matrix $\pi = \{\pi_D(E)\}_{D \subseteq V, E \subseteq U}$ where, for every $E \subseteq U$, the column vector $\pi(E)$ is the probability distribution of $Y_V$ given $(X_E = 1_E, X_{U \setminus E} = 0_{U \setminus E})$ and, more specifically, $\pi_D(E) = \mathrm{pr}(Y_D = 1_D, Y_{V \setminus D} = 0_{V \setminus D} \mid X_E = 1_E, X_{U \setminus E} = 0_{U \setminus E})$. We assume that all the entries of $\pi$ are strictly positive. In the following, for a subset $D \subseteq V$ we use the suppressed notation $Y_D = 1$ to denote $Y_D = 1_D$ and similarly for $Y_D = 0$ and the subvectors of $X_U$. Given three random vectors $X$, $Y$ and $Z$, we write $X \perp\!\!\!\perp Y|Z$ to say that $X$ is independent of $Y$ given $Z$ (Dawid, 1979) or, in the case where $Y$ and $Z$ are not random, that the conditional distribution of $X$ given $Y$ and $Z$ does not depend on $Y$.

In regression models for categorical responses a linear regression is typically related to the response variables via a link function $\theta(\pi)$. All the link functions considered in this paper are such that, for every $E \subseteq U$, the vector $\theta(E) = \theta(\pi(E))$ parameterizes the distribution of $Y_V|(X_E = 1, X_{U \setminus E} = 0)$. In this way, the link function induces a matrix $\theta = \{\theta_D(E)\}_{D \subseteq V, E \subseteq U}$ and the associated linear regression, in the saturated case, has form

$$\theta_D(E) = \sum_{E' \subseteq E} \beta_D^{\langle \theta \rangle}(E') \qquad \text{for every } D \subseteq V, E \subseteq U \tag{3}$$

where $\beta^{\langle \theta \rangle} = \{\beta_D^{\langle \theta \rangle}(E)\}_{D \subseteq V, E \subseteq U}$ is a matrix of regression coefficients. It follows from (1) and (2) that (3) can be written in matrix form as

$$\theta = \beta^{\langle \theta \rangle} Z_U \qquad \text{so that} \qquad \beta^{\langle \theta \rangle} = \theta M_U. \tag{4}$$

The regression setting (3) involves multivariate combinations of both the responses and the covariates and for this reason it is important to explicitly distinguish between the *D-response associations* which are given by $\theta_D$ for $D \subseteq V$ with $|D| > 1$, and the

*E-covariate interactions* given by $\beta^{\langle\theta\rangle}(E)$ for $E \subseteq U$ with $|E| > 1$. Hence, $\beta_D^{\langle\theta\rangle}(E)$ encodes the effect of the $E$-covariate interaction on the $D$-response association, given a fixed level of $X_{U\setminus E}$ that, without loss of generality, in our approach is the zero level.

**Example 3.2 (Matrix notation continued.)** *In all the examples of this paper we use the variables of the multimorbidity data as given in Section 5. Specifically, we consider three of the response variables, which are* $Y_b =$ Bone fracture, $Y_c =$ Cardiovascular disease *and* $Y_d =$ Diabetes, *with level 1 encoding the presence of the disease, and the most relevant covariates, that is,* $X_h =$ HIV *with the level 1 encoding the presence of the infection and* $X_a =$ Age *with the value 1 for patients aged 45 or more. Hence, if the link function is the matrix* $\theta$ *given in Example 3.1, then the matrix of regression coefficients* $\beta^{\langle\theta\rangle}$ *is*

$$\beta^{\langle\theta\rangle} = \begin{pmatrix} \beta_\emptyset^{\langle\theta\rangle}(\emptyset) & \beta_\emptyset^{\langle\theta\rangle}(h) & \beta_\emptyset^{\langle\theta\rangle}(a) & \beta_\emptyset^{\langle\theta\rangle}(ha) \\ \beta_b^{\langle\theta\rangle}(\emptyset) & \beta_b^{\langle\theta\rangle}(h) & \beta_b^{\langle\theta\rangle}(a) & \beta_b^{\langle\theta\rangle}(ha) \\ \beta_c^{\langle\theta\rangle}(\emptyset) & \beta_c^{\langle\theta\rangle}(h) & \beta_c^{\langle\theta\rangle}(a) & \beta_c^{\langle\theta\rangle}(ha) \\ \beta_d^{\langle\theta\rangle}(\emptyset) & \beta_d^{\langle\theta\rangle}(h) & \beta_d^{\langle\theta\rangle}(a) & \beta_d^{\langle\theta\rangle}(ha) \\ \beta_{bc}^{\langle\theta\rangle}(\emptyset) & \beta_{bc}^{\langle\theta\rangle}(h) & \beta_{bc}^{\langle\theta\rangle}(a) & \beta_{bc}^{\langle\theta\rangle}(ha) \\ \beta_{bd}^{\langle\theta\rangle}(\emptyset) & \beta_{bd}^{\langle\theta\rangle}(h) & \beta_{bd}^{\langle\theta\rangle}(a) & \beta_{bd}^{\langle\theta\rangle}(ha) \\ \beta_{cd}^{\langle\theta\rangle}(\emptyset) & \beta_{cd}^{\langle\theta\rangle}(h) & \beta_{cd}^{\langle\theta\rangle}(a) & \beta_{cd}^{\langle\theta\rangle}(ha) \\ \beta_{bcd}^{\langle\theta\rangle}(\emptyset) & \beta_{bcd}^{\langle\theta\rangle}(h) & \beta_{bcd}^{\langle\theta\rangle}(a) & \beta_{bcd}^{\langle\theta\rangle}(ha) \end{pmatrix} = \begin{pmatrix} \beta_\emptyset^{\langle\theta\rangle} \\ \beta_b^{\langle\theta\rangle} \\ \beta_c^{\langle\theta\rangle} \\ \beta_d^{\langle\theta\rangle} \\ \beta_{bc}^{\langle\theta\rangle} \\ \beta_{bd}^{\langle\theta\rangle} \\ \beta_{cd}^{\langle\theta\rangle} \\ \beta_{bcd}^{\langle\theta\rangle} \end{pmatrix}$$

*Equation (3) states that there exists a one-to-one Möbius inversion relationship between every row* $\theta_D$ *of* $\theta$ *and the corresponding row* $\beta_D^{\langle\theta\rangle}$ *of* $\beta^{\langle\theta\rangle}$. *For instance,* $\theta_b$ *is the link function of the marginal distribution of bone fracture, and each of its entries,* $\theta_b(E)$ *for* $E \subseteq U$, *can be computed by considering* $\beta_b^{\langle\theta\rangle}$ *and then by taking the sum of its entries which are indexed by a subset of* $E$. *More specifically, for* $E = \{h\}$ *it holds that* $\theta_b(h) = \beta_b^{\langle\theta\rangle}(\emptyset) + \beta_b^{\langle\theta\rangle}(h)$ *whereas for* $E = \{h, a\}$ *it holds that* $\theta_b(ha) = \beta_b^{\langle\theta\rangle}(\emptyset) + \beta_b^{\langle\theta\rangle}(h) + \beta_b^{\langle\theta\rangle}(a) + \beta_b^{\langle\theta\rangle}(ha)$. *Hence,* $\beta_b^{\langle\theta\rangle}(h)$ *and* $\beta_b^{\langle\theta\rangle}(ha)$ *encode the effect of HIV and of the interaction of HIV and age, respectively, on bone fracture. Similarly,* $\theta_{bd}$ *is the response association of bone fracture and diabetes. Hence, for* $E = \{h\}$ *it holds that* $\theta_{bd}(h) = \beta_{bd}^{\langle\theta\rangle}(\emptyset) + \beta_{bd}^{\langle\theta\rangle}(h)$ *whereas for* $E = \{h, a\}$ *it holds that* $\theta_{bd}(ha) = \beta_{bd}^{\langle\theta\rangle}(\emptyset) + \beta_{bd}^{\langle\theta\rangle}(h) + \beta_{bd}^{\langle\theta\rangle}(a) + \beta_{bd}^{\langle\theta\rangle}(ha)$. *Here,* $\beta_{bd}^{\langle\theta\rangle}(h)$ *and* $\beta_{bd}^{\langle\theta\rangle}(ha)$ *encode the effect of HIV and of the interaction of HIV and age, respectively, on the* $\{b, d\}$*-response association* $\theta_{bd}$.

There exists an extensive literature on models for categorical data analysis but, remarkably, Lang (1996), extending previous work by Lang and Agresti (1994), introduced a very general method to specify regression models for categorical data thereby defining an extremely broad class of models named *generalized log-linear models*. This includes, as special cases, many of the existing models for multiple categorical responses such as log-linear and, more generally, marginal log-linear models (Bergsma et al., 2009). In particular, in marginal regression modelling a relevant instance within this class is obtained when $\theta(\cdot)$ is the multivariate logistic link function denoted by $\eta(\cdot)$ (McCullagh and Nelder, 1989; Glonek and McCullagh, 1995; Molenberghs and Lesaffre, 1999; Bartolucci et al., 2007; Bergsma et al., 2009; Marchetti and Lupparelli, 2011). This induces the parameterization $\eta = \{\eta_D(E)\}_{D \subseteq V, E \subseteq U}$ where $\eta_D(E)$ is the $|D|$-way log-linear interaction computed in the margin $Y_D|(X_E = 1, X_{U \setminus E} = 0)$ and, more specifically, $\eta_{\{v\}}$ is the usual logistic link for every $v \in V$. Marchetti and Lupparelli (2011) considered the regression framework (3) with $\theta = \eta$ and showed that $\beta_D^{\langle \eta \rangle}(E)$ is the $|D \cup E|$-way log-linear interaction computed in the distributions of $(Y_D, X_U)$.

**Example 3.3 (Multivariate logistic regression)** *Consider the case $V = \{b, c\}$ and $U = \{h\}$. The multivariate logistic parameters are*

$$\eta_i(\emptyset) = \log \frac{\mathrm{pr}(Y_i = 1|X_h = 0)}{1 - \mathrm{pr}(Y_i = 1|X_h = 0)} \qquad \textit{and} \qquad \eta_i(h) = \log \frac{\mathrm{pr}(Y_i = 1|X_h = 1)}{1 - \mathrm{pr}(Y_i = 1|X_h = 1)}$$

*for $i \in \{b, c\}$, which are logit links. Furthermore, if we denote by $OR(Y_{bc}|\emptyset)$ and by $OR(Y_{bc}|h)$ the odds ratio between bone fracture and cardiovascular disease for HIV-negative and HIV-positive patients, respectively, then*

$$\eta_{bc}(\emptyset) = \log OR(Y_{bc}|\emptyset) \qquad \textit{and} \qquad \eta_{bc}(h) = \log OR(Y_{bc}|h)$$

*so that the coefficients for the regression of $\eta_{bc}$ are*

$$\beta_{bc}^{\langle \eta \rangle}(\emptyset) = \log OR(Y_{bc}|\emptyset) \qquad \textit{and} \qquad \beta_{bc}^{\langle \eta \rangle}(h) = \log \frac{OR(Y_{bc}|h)}{OR(Y_{bc}|\emptyset)}.$$

We now move from the saturated model to submodels defined by means of linear constraints on the regression coefficients and, more specifically, to submodels with regression coefficients equal to zero. If for every subset of $U$, denoted as $E$, that has non-empty intersection with $U' \subseteq U$, it holds that $\beta_D^{\langle \theta \rangle}(E) = 0$, then the covariates $X_{U'}$ have no-effect on $\theta_D$. The following lemma states a connection between no-effect

of a subset of variables and linear constraints in $\theta$.

**Lemma 3.1** *Let $\theta = \{\theta_D(E)\}_{D\subseteq V, E\subseteq U}$ be a real matrix with entries indexed by two nonempty sets $V$ and $U$. If $\beta^{\langle\theta\rangle} = \theta M_U$ and $U' \subseteq U$ then the following are equivalent, for every $D \subseteq V$, as they both say that $X_{U'}$ has no effect on $\theta_D$;*

*(i) $\beta_D^{\langle\theta\rangle}(E) = 0$ for every $E \subseteq U$ such that $E \cap U' \neq \emptyset$;*

*(ii) $\theta_D(E_1) = \theta_D(E_2)$ for every $E_1, E_2 \subseteq U$ such that $E_1 \backslash U' = E_2 \backslash U'$.*

*Proof.* See Appendix A. □

# 4 Log-mean and log-mean linear regression models

The multivariate Bernoulli distribution belongs to the natural exponential family, and the mean parameter associated with the distribution $\pi(E)$ is $\mu(E)$ where $\mu_D(E) = \mathrm{pr}(Y_D = 1 | X_E = 1, X_{U\backslash E} = 0)$ for $D \subseteq V$. Drton and Richardson (2008) used the mean parameter to parameterize graphical models of marginal independence and called it the Möbius parameter because $\mu(E) = Z_V \pi(E)$ for every $E \subseteq V$. Subsequently, Drton (2009) used the matrix $\mu = Z_V \pi$ to parameterize regression graph models; see also Roverato et al. (2013) and Roverato (2015). The mean parameterization has the disadvantage that submodels of interest are defined by, non-linear, multiplicative constraints. Roverato et al. (2013) introduced the log-mean linear parameter $\gamma$ defined as a log-linear expansion of the mean parameters, formally,

$$\gamma = M_V^\top \log(\mu), \tag{5}$$

and showed that this approach improves on the mean parameterization as submodels of interest can be specified by setting certain zero log-mean linear interactions. We remark that both the mean and the log-mean linear parameters are not variation independent, in the sense that setting some parameters to particular values may restrict the valid range of other parameters. As a consequence, unlike variation independent parameterizations, it might be more difficult to interpret separately each parameter.

Here, we show that the analysis of multimorbidity data can be effectively approached by applying the theory described in the previous section to the log-mean (LM) and the log-mean linear (LML) parameterizations to develop the LM and the LML regression model, respectively.

## 4.1 Log-mean regression

The LM regression model is obtained by setting $\theta$ in (3) equal to the logarithm of the mean parameter, $\log(\mu)$, so that, in the saturated case, $\log(\mu) = \beta^{\langle\mu\rangle} Z_U$ that, in its extended form, is

$$\log \mu_D(E) = \sum_{E' \subseteq E} \beta_D^{\langle\mu\rangle}(E') \quad \text{for every } D \subseteq V, E \subseteq U \tag{6}$$

where for $D = \emptyset$ the equation (6) is trivial for every $E \subseteq U$ because $\mu_\emptyset$ is a row vector of ones so that both $\log(\mu_\emptyset) = 0$ and $\beta_\emptyset^{\langle\mu\rangle} = \log(\mu_\emptyset) M_U = 0$.

LM regression can be regarded as one of the possible alternative ways to parameterize the distribution of $Y_V | X_U$, but it is of special interest for the application considered in this paper. This can be seen by noticing that one can write $\mu_D(E) = \text{pr}(Y^D = 1 | X_E = 1, X_{U \setminus E} = 0)$ where $Y^D = \prod_{v \in D} Y_v$ is the binary random variable associated with the multimorbidity pattern $D$. More specifically, $\text{pr}(Y^D = 1)$ is the probability that the multimorbidity pattern $D$ occurs, and therefore (6) is a sequence of regressions corresponding to the univariate binary responses $Y^D$ for $\emptyset \neq D \subseteq V$.

Similarly to Poisson regression, the parameters of LM regression can be interpreted in terms of relative risks. We first consider the case $U = \{u\}$, so that $|U| = 1$. Hence, if we denote the relative risk of an event $\mathcal{E}$ with respect to the two groups identified by $X_u = 1$ and $X_u = 0$ by

$$RR_u(\mathcal{E}) = \frac{\text{pr}(\mathcal{E} | X_u = 1)}{\text{pr}(\mathcal{E} | X_u = 0)} \tag{7}$$

then for every $v \in V$ it holds that $\beta_v^{\langle\mu\rangle}(u) = \log RR_u(Y_v = 1)$ whereas, more generally, for $D \subseteq V$

$$\beta_D^{\langle\mu\rangle}(u) = \log RR_u(Y^D = 1) = \log RR_u(\cap_{v \in D} \{Y_v = 1\}) \tag{8}$$

where we use the convention that $RR_u(Y^\emptyset = 1) = 1$.

**Example 4.1 (LM regression)** *Consider the case where $V = \{b, c, d\}$ and $U = \{h\}$. Then the saturated LM regression model contains a regression equation for every*

*$D \subseteq V$, where the case $D = \emptyset$ is trivial. For $|D| = 1$, equation (6) has form*

$$\begin{aligned} \log \mu_i(\emptyset) = \log \mathrm{pr}(Y_i = 1 | X_h = 0) &= \beta_i^{\langle \mu \rangle}(\emptyset) \\ \log \mu_i(h) = \log \mathrm{pr}(Y_i = 1 | X_h = 1) &= \beta_i^{\langle \mu \rangle}(\emptyset) + \beta_i^{\langle \mu \rangle}(h) \end{aligned}$$

*where, for every $i \in V$, $\beta_i^{\langle \mu \rangle}(h) = \log RR_h(Y_i = 1)$ that is the log-relative risk of the disease $Y_i$ for HIV-positive patients compared to HIV-negative patients. For $|D| = 2$,*

$$\begin{aligned} \log \mu_{ij}(\emptyset) = \log \mathrm{pr}(Y_i = 1, Y_j = 1 | X_h = 0) &= \beta_{ij}^{\langle \mu \rangle}(\emptyset) \\ \log \mu_{ij}(h) = \log \mathrm{pr}(Y_i = 1, Y_j = 1 | X_h = 1) &= \beta_{ij}^{\langle \mu \rangle}(\emptyset) + \beta_{ij}^{\langle \mu \rangle}(h) \end{aligned}$$

*where, for every $i, j \in V$ with $i \neq j$, $\beta_{ij}^{\langle \mu \rangle}(h) = \log RR_h(Y^{ij} = 1)$ that is the log-relative risk of the co-occurrence of the diseases $Y_i$ and $Y_j$ for HIV-positive patients compared to HIV-negative patients. Finally, for $|D| = 3$ so that $D = V$,*

$$\begin{aligned} \log \mu_{bcd}(\emptyset) = \log \mathrm{pr}(Y_b = 1, Y_c = 1, Y_d = 1 | X_h = 0) &= \beta_{bcd}^{\langle \mu \rangle}(\emptyset) \\ \log \mu_{bcd}(h) = \log \mathrm{pr}(Y_b = 1, Y_c = 1, Y_d = 1 | X_h = 1) &= \beta_{bcd}^{\langle \mu \rangle}(\emptyset) + \beta_{bcd}^{\langle \mu \rangle}(h) \end{aligned}$$

*where $\beta_{bcd}^{\langle \mu \rangle}(h) = \log RR_h(Y^{bcd} = 1)$ is the log-relative risk of the co-occurrence of the three diseases for HIV-positive patients compared to HIV-negative patients.*

Consider the regression equation relative to the subset $\emptyset \neq D \subseteq V$. It follows from (6) that if for an element $u \in U$ it holds that $\beta_D^{\langle \mu \rangle}(E) = 0$ for every $E \subseteq U$ such that $u \in E$, then $Y^D \perp\!\!\!\perp X_u | X_{U \setminus \{u\}}$. This can be easily extended to a subset of covariates $X_{U'}$ with $\emptyset \neq U' \subseteq U$, whereas to generalize this result to the vector $Y_D$ it is necessary to consider a collection of regression equations as shown below.

**Proposition 4.1** *Let $Y_D$ be the subvector of $Y_V$ indexed by $\emptyset \neq D \subseteq V$. For a subset $\emptyset \neq U' \subseteq U$, it holds that $Y_D \perp\!\!\!\perp X_{U'} | X_{U \setminus U'}$ if and only if $\beta_{D'}^{\langle \mu \rangle}(E) = 0$ for every $D' \subseteq D$ and $E \subseteq U$ such that $E \cap U' \neq \emptyset$.*

*Proof.* See Appendix A. □

We can conclude that it makes sense to focus on submodels characterized by zero LM regression coefficients because they encode interpretable relationships, possibly implying that one or more covariates have no-effect on the distribution of $Y^D$, or even on the joint distribution of $Y_D$. However, this approach did not identify any missing effects in the application to multimorbidity data. One reason for that is that the zero

pattern of regression coefficients described in Proposition 4.1 implies no-effect of $X_{U'}$ on $Y_D$ and therefore that $X_{U'}$ has no-effect on $Y_v$ for every single $v \in D$. Consider the case $|U| = 1$ of a unique covariate $X_h$ representing the HIV-infection. It is well established that HIV has a relevant effect on each of the comorbidites considered singularly, that is, $\beta_v^{\langle\mu\rangle}(h) \neq 0$ for every $v \in V$. Although in principle it is possible for a covariate $X_h$ with a non-zero effect on $Y_v$ for some $v \in D$ to have coefficients equal to zero in the regression relative to $Y^D$, this hardly happens in practice because of the strong association existing between $Y^D$ and every $Y_v$ with $v \in D$; recall that $Y_v = 0$ for any $v \in D$ implies $Y_D = 0$ and, conversely, $Y^D = 1$ implies $Y_v = 1$ for every $v \in D$. In other words, it is well known that, if $X_h = 1$ for HIV-postive patients, then $\text{pr}(Y_v = 1|X_h = 1) > \text{pr}(Y_v = 1|X_h = 0)$ for every comorbidity $v \in V$, and therefore it is reasonable to expect that also $\text{pr}(Y^D = 1|X_h = 1) > \text{pr}(Y^D = 1|X_h = 0)$ for every comorbidity pattern $D \subseteq V$, with $|D| > 1$.

To disclose the usefulness of LM regression for the application considered in this paper, it is necessary to propose a different approach. In the multimorbidity analysis, it is useful to distinguish between two different effects of HIV, specifically, one can be interest in the effect of HIV (i) on the prevalence of a comorbidity pattern $D$ and (ii) on the association among comorbidities in $D$. As discussed above, the former is in general a well known matter as multimorbidity shows a higher prevalence in infected patients. Indeed, the main question is whether HIV plays a role in the way single comorbidities combine together to produce the comorbidity pattern $D$, that is in the association of variables in $Y_D$. We approach the problem by considering the extreme case of no-effect of HIV on the $D$-response associations because responses are conditionally independent given the covariates. Our standpoint is that if for a subset $D \subseteq V$ there exists a proper partition $A \cup B = D$ such that $Y_A \perp\!\!\!\perp Y_B | X_U$, then there is no-effect of the covariates on the $|D|$-way association of $Y_D$. This allows us to compare the performance of different link functions in multimorbidity analysis and, more importantly, to provide a clear interpretation to the effect of HIV in LM regression. In the Example 4.2 below we illustrate how this idea can be formalized.

**Example 4.2 (LM regression vs. multivariate logistic regression)** *For the case $V = \{b, d\}$ and $U = \{h\}$, assume $Y_b \perp\!\!\!\perp Y_d | X_h$ so that we can say that there is no-effect of $X_h = HIV$ on the association of $Y_b$ and $Y_d$. To see this in practice it is sufficient to consider any measure of association that represents independence by a constant value, for instance the value zero, so that the association is the same for the two values of $X_h$. Exploiting the factorization of the probability function of $Y_{bd}|X_h$ implied by conditional independence, it is not difficult to see that under the LM regression*

*both*

$$Y_b \perp\!\!\!\perp Y_d | X_h \quad \Longleftrightarrow \quad \beta_{bd}^{\langle\mu\rangle}(\emptyset) = \beta_b^{\langle\mu\rangle}(\emptyset) + \beta_d^{\langle\mu\rangle}(\emptyset) \quad \textit{and} \quad \beta_{bd}^{\langle\mu\rangle}(u) = \beta_b^{\langle\mu\rangle}(h) + \beta_d^{\langle\mu\rangle}(h),$$

*and, under the multivariate logistic regression,*

$$Y_b \perp\!\!\!\perp Y_d | X_h \quad \Longleftrightarrow \quad \beta_{bd}^{\langle\eta\rangle}(\emptyset) = 0 \quad \textit{and} \quad \beta_{bd}^{\langle\eta\rangle}(h) = 0.$$

*The effect of HIV on the joint distribution of $Y_{bd}$ is represented by $\beta_{bd}^{\langle\mu\rangle}(h)$ in LM regression and by $\beta_{bd}^{\langle\eta\rangle}(h)$ in multivariate logistic regression and therefore both*

$$\beta_{bd}^{\langle\mu\rangle}(h) = \beta_b^{\langle\mu\rangle}(h) + \beta_d^{\langle\mu\rangle}(h) \tag{9}$$

*and*

$$\beta_{bd}^{\langle\eta\rangle}(h) = 0 \tag{10}$$

*can be used to state that there is no-effect of HIV on the association of $Y_b$ and $Y_d$. However, (9) and (10) refer to different kinds of association because, although they are necessary conditions for $Y_b \perp\!\!\!\perp Y_d | X_h$, they are not sufficient for the same condition to hold true, and it can be easily checked that neither (9) implies (10) nor (10) implies (9). From this perspective, it makes little sense to state that* HIV has no-effect on the association of $Y_b$ and $Y_d$ *if a clear interpretation to equalities (9) and (10) is not provided.*

*The interpretation of equality (9) is straightforward. It states a connection between regressions of different order and implies that the effect of HIV on the distribution of $Y^{bd}$ is explained by the effects of HIV on the marginal distributions of $Y_b$ and $Y_d$. More interestingly, it provides an useful insight on the behaviour of relative risks because (9) is equivalent to*

$$RR_h(\{Y_b = 1\} \cap \{Y_d = 1\}) = RR_h(Y_b = 1) \times RR_h(Y_d = 1) \tag{11}$$

*that is, if we are willing to interpret the effect of HIV by means of relative risks, then (9) allows one to carry out the analysis marginally on the regression equations for the main responses, because the computation of the relative risk of the joint event $\{Y_b = 1\} \cap \{Y_d = 1\}$ does not require the joint distribution of $Y_{bd}|X_h$ but only the marginal distributions of $Y_b|X_h$ and $Y_d|X_h$ as in the case where $Y_b \perp\!\!\!\perp Y_d | X_h$.*

*Multivariate logistic regression parameters are naturally associated with odds ra-*

*tios, which play a very fundamental role among association measures for categorical data. However, we deem that, in this context, the interpretation of (9) is more straightforward than (10). Indeed, the latter is equivalent to the identity* $OR(Y_{bd}|\emptyset) = OR(Y_{bd}|h)$ *that, unlike (11), does not explains how the marginal effects of HIV on bone fracture and diabetes combine together to give the effect of HIV on the* $\{b, d\}$*-response association.*

This example shows that the LM regression coefficients provide a clear way to investigate the effect of a covariate on the association of two binary variables and the following theorem generalises this result to $D$-response associations.

**Theorem 4.2** *Let* $\mu$ *be the mean parameter of* $Y_V|X_U$ *and* $\beta^{\langle\mu\rangle} = \log(\mu)M_U$*. Then, for a pair of disjoint nonempty subsets* $A$ *and* $B$ *of* $V$ *it holds that* $Y_A \perp\!\!\!\perp Y_B|X_U$ *if and only if for every* $D \subseteq A \cup B$ *and* $E \subseteq U$*, with both* $D \cap A \neq \emptyset$ *and* $D \cap B \neq \emptyset$*, it holds that* $\beta_D^{\langle\mu\rangle}(E) = \mathcal{B}_D^{\langle\mu\rangle}(E)$ *where*

$$\mathcal{B}_D^{\langle\mu\rangle}(E) = -\sum_{D' \subset D} (-1)^{|D \setminus D'|}\, \beta_{D'}^{\langle\mu\rangle}(E). \tag{12}$$

*Proof.* See Appendix A. □

Theorem 4.2 shows that, whenever $Y_D$ can be split into two conditionally independent subvectors $Y_A$ and $Y_B$, then every coefficient of the LM regression with response $\log(\mu_D)$ can be written as a linear combination of the corresponding coefficients in lower-order regressions. Hence, the difference between case and control patients with respect to $Y^D$ is not given by the effect of HIV on the association of all the comorbidities in $D$, but it is only the consequence of the effect that HIV has on the occurrence of the subsets $A$ and $B$ of comorbidity patterns.

For the case $U = \{u\}$, the relationship $\beta_D^{\langle\mu\rangle}(u) = \mathcal{B}_D^{\langle\mu\rangle}(u)$ can be used to state that $X_u$ has no-effect on the $D$-response association and, as well as in Example 4.2, this statement means that the effect of $X_u$ on the distribution of $Y^D$ is explained by the corresponding coefficients in lower-order regressions. Furtheremore, it follows immediately from (8) that the equality $\beta_D^{\langle\mu\rangle}(u) = \mathcal{B}_D^{\langle\mu\rangle}(u)$ can be equivalently stated in terms of factorization of the relative risk $RR_u(Y^D = 1)$ with respect to the collection of relative risks $RR_u(Y^{D'} = 1)$ for $D' \subset D$.

**Example 4.3 (LM regression cont.)** *For the LM regression in Example 4.1 consider the case where* $Y_A \perp\!\!\!\perp Y_B|X_h$ *where* $A$ *and* $B$ *are nonempty, disjoint subsets of* $V$

*such that $A \cup B = V$. Then, it follows from Theorem 4.2 both that*

$$\beta_{bcd}^{\langle\mu\rangle}(\emptyset) = -\beta_{b}^{\langle\mu\rangle}(\emptyset) - \beta_{c}^{\langle\mu\rangle}(\emptyset) - \beta_{d}^{\langle\mu\rangle}(\emptyset) + \beta_{bc}^{\langle\mu\rangle}(\emptyset) + \beta_{bd}^{\langle\mu\rangle}(\emptyset) + \beta_{cd}^{\langle\mu\rangle}(\emptyset)$$

*and that*

$$\beta_{bcd}^{\langle\mu\rangle}(h) = -\beta_{b}^{\langle\mu\rangle}(h) - \beta_{c}^{\langle\mu\rangle}(h) - \beta_{d}^{\langle\mu\rangle}(h) + \beta_{bc}^{\langle\mu\rangle}(h) + \beta_{bd}^{\langle\mu\rangle}(h) + \beta_{cd}^{\langle\mu\rangle}(h)$$
$$\Updownarrow$$
$$RR_h(Y^{bcd}=1) = \frac{RR_h(Y^{bc}=1)\times RR_h(Y^{bd}=1)\times RR_h(Y^{cd}=1)}{RR_h(Y^{b}=1)\times RR_h(Y^{c}=1)\times RR_h(Y^{d}=1)}.$$

More generally, for the case $|U| > 1$ it follows from Theorem 4.2 that the relationship

$$\beta_{D}^{\langle\mu\rangle}(E) = \mathcal{B}_{D}^{\langle\mu\rangle}(E) \quad \text{for every } E \subseteq U \text{ such that } u \in E \tag{13}$$

implies that every regression coefficient involving $X_u$ is the linear combination of the corresponding coefficients of lower-order regressions, and can be used to state that $X_u$ has no-effect on the $D$-response association. Also in this case, (13) can be interpreted in terms of relative risks. Consider the collection of relative risks of the event $Y^D = 1$, with respect to $X_u$, conditionally on the values of the remaining covariates $X_{U\setminus\{u\}}$, given by

$$RR_u(Y^D = 1|E) = \frac{\text{pr}(Y^D = 1|X_u = 1,\ X_E = 1, X_{U\setminus(E\cup\{u\})} = 0)}{\text{pr}(Y^D = 1|X_u = 0,\ X_E = 1, X_{U\setminus(E\cup\{u\})} = 0)} \quad \text{for } E \subseteq U\setminus\{u\}; \tag{14}$$

we recall that we set $RR_u(Y^{\emptyset} = 1|E) = 1$. Then we associate to every element of (14) a reference value defined as follows.

**Definition 4.1** *For $D \subseteq V$ with $|D| > 1$, the* reference relative risk *of the event $Y^D = 1$ with respect to $X_u$ and a subset $E \subseteq U\setminus\{u\}$ is defined as*

$$\mathcal{RR}_u(Y^D = 1|E) = \prod_{D'\subset D} RR_u(Y^{D'} = 1|E)^{(-1)^{|D\setminus D'|+1}}.$$

The following lemma states the connection between relative risks and regression coefficients as well as between reference relative risks in Definition 4.1 and Theorem 4.2.

**Lemma 4.3** *Let $\mu$ be the mean parameter of $Y_V|X_U$ and $\beta^{\langle\mu\rangle} = \log(\mu)M_U$. Then,*

*for every $D \subseteq V$, $u \in U$ and $E \subseteq U \backslash \{u\}$ it holds that*

$$\log RR_u(Y^D = 1|E) = \sum_{E' \subseteq E} \beta_D^{\langle \mu \rangle}(E' \cup \{u\})$$

*and, for $|D| > 1$, that*

$$\log \mathcal{RR}_u(Y^D = 1|E) = \sum_{E' \subseteq E} \mathcal{B}_D^{\langle \mu \rangle}(E' \cup \{u\}).$$

*Proof.* See Appendix A. □

The introduction of the reference relative risks is motivated by the following result,

**Corollary 4.4** *Let the subset $D \subseteq V$ be such that there exists a proper partition $D = A \cup B$, with $A, B \neq \emptyset$, satisfying $Y_A \perp\!\!\!\perp Y_B | X_U$. Then for every $u \in U$ it holds that*

$$RR_u(Y^D = 1|E) = \mathcal{RR}_u(Y^D = 1|E) \quad \textit{for every } E \subseteq U \backslash \{u\}. \tag{15}$$

*Proof.* See Appendix A. □

Hence $\mathcal{RR}_u(Y^D = 1|E)$ is a reference value in the sense that it is the value taken by the corresponding relative risk when $Y_D$ can be split into two conditionally independent subvectors. If (15) is satisfied, then all the conditional relative risks of the events $\{Y^D = 1\}$, with respect to $X_u$, are equal to their reference value. This implies that, as far as the relative risk of the multimorbidity pattern $D$ is of concern, the analysis can be carried out marginally on the distributions of $Y_{D'}|X_U$ with $D' \subset D$. The following corollary shows that one can interpret (13) in terms of relative risks because it is equivalent to (15).

**Corollary 4.5** *Let $\mu$ be the mean parameter of $Y_V|X_U$ and $\beta^{\langle \mu \rangle} = \log(\mu) M_U$. For any $D \subseteq V$, with $|D| > 1$, and $u \in U$ the relationship (13) is satisfied if and only if (15) is satisfied.*

*Proof.* See Appendix A. □

We can conclude that LM regression provides a useful framework to investigate the effect of covariates on the association of responses. Submodels of interest involve possibly zero regression coefficients, as in Proposition 4.1, but, more commonly, regression

coefficients which are linear combination of lower-order coefficients as in Theorem 4.2. This approach can be easily implemented because (i) it involves submodels defined by linear constraints on the regression parameters and (ii) the computation of $\mathcal{B}_D^{\langle\mu\rangle}(E)$ in (12) does not require to specify explicitly the partition of $Y_D$ into independent subvectors. A shortcoming of LM regression is that in order to interpret the values of regression coefficients in terms of conditional independence the coefficients must be contrasted with the theoretical values (12) given in Theorem 4.2. In the next section we show that LML regression provides a solution to this problem.

## 4.2 Log-mean linear regression

The LML regression model is obtained by setting $\theta$ in (3) equal to the LML parameter $\gamma = M_V^\top \log(\mu)$ in (5), so that, in the saturated case, it follows from (4) that $\beta^{\langle\gamma\rangle} = \gamma M_U$ and $\gamma = \beta^{\langle\gamma\rangle} Z_U$; the latter can be written as

$$\gamma_D(E) = \sum_{E' \subseteq E} \beta_D^{\langle\gamma\rangle}(E') \quad \text{for every } D \subseteq V, E \subseteq U. \tag{16}$$

There is a close connection between LM and LML regression given by a linear relationship between $\beta^{\langle\gamma\rangle}$ and $\beta^{\langle\mu\rangle}$.

**Lemma 4.6** *Let $\mu$ and $\gamma$ be the mean and LML parameter, respectively, of $Y_V|X_U$ so that $\beta^{\langle\mu\rangle} = \log(\mu)M_U$ and $\beta^{\langle\gamma\rangle} = \gamma M_U$. Then it holds that $\beta^{\langle\gamma\rangle} = M_V^\top \beta^{\langle\mu\rangle}$, that is for every $D \subseteq V$ and $E \subseteq U$*

$$\beta_D^{\langle\gamma\rangle}(E) = \sum_{D' \subseteq D} (-1)^{|D \setminus D'|}\, \beta_{D'}^{\langle\mu\rangle}(E). \tag{17}$$

*Proof.* See Appendix A. □

As a consequence of Lemma 4.6, submodels defined by zero LM regression coefficients as in Proposition 4.1 can be equivalently stated by setting to zero the corresponding LML regression coefficients.

**Corollary 4.7** *Let $Y_D$ be the subvector of $Y_V$ indexed by $\emptyset \neq D \subseteq V$. For a subset $\emptyset \neq U' \subseteq U$, it holds that $Y_D \perp\!\!\!\perp X_{U'}|X_{U\setminus U'}$ if and only if $\beta_{D'}^{\langle\gamma\rangle}(E) = 0$ for every $D' \subseteq D$ and $E \subseteq U$ such that $E \cap U' \neq \emptyset$.*

*Proof.* See Appendix A. □

Furthermore, it follows immediately from (17) that for $|D| \leq 1$ it holds that $\beta_D^{\langle\gamma\rangle} = \beta_D^{\langle\mu\rangle}$ whereas for $|D| > 1$ LML regression coefficients allow one to immediately check whether a LM regression coefficients coincides with the associated theoretical value given in Theorem 4.2.

**Proposition 4.8** *Let $\mu$ and $\gamma$ be the mean and LML parameter, respectively, of $Y_V|X_U$ so that $\beta^{\langle\mu\rangle} = \log(\mu)M_U$ and $\beta^{\langle\gamma\rangle} = \gamma M_U$. Then for every $D \subseteq V$, such that $|D| > 1$, and $E \subseteq U$ it holds that $\beta_D^{\langle\gamma\rangle}(E) = \beta_D^{\langle\mu\rangle}(E) - \mathcal{B}_D^{\langle\mu\rangle}(E)$ so that*

$$\beta_D^{\langle\gamma\rangle}(E) = 0 \quad \textit{if and only if} \quad \beta_D^{\langle\mu\rangle}(E) = \mathcal{B}_D^{\langle\mu\rangle}(E).$$

*Proof.* This is an immediate consequence of (17). □

Hence, the relationship

$$\beta_D^{\langle\gamma\rangle}(E) = 0 \quad \text{for every } E \subseteq U \text{ such that } u \in E \tag{18}$$

is equivalent to both (13) and (15) and we can conclude that LML regression is equivalent to LM regression for the purposes of our analysis, with the advantage that all the submodels of interest can be specified by setting LML regression coefficients to zero. Furthermore, the value of the regression coefficients $\beta^{\langle\gamma\rangle}$ can be used to contrast relative risks with the corresponding reference values as follows.

**Corollary 4.9** *Let $\gamma$ be the LML parameter of $Y_V|X_U$ so that $\beta^{\langle\gamma\rangle} = \gamma M_U$. Then for every $D \subseteq V$, such that $|D| > 1$, and $E \subseteq U$ it holds that*

$$\log \frac{RR_u(Y^D = 1|E)}{\mathcal{RR}_u(Y^D = 1|E)} = \sum_{E' \subseteq E} \beta_D^{\langle\gamma\rangle}(E' \cup \{u\}) \tag{19}$$

*Proof.* See Appendix A. □

Note that, interestingly, Corollary 4.9 implies that for the case $|U| = 1$ the LML regression coefficients such that $|D| > 1$ have an immediate interpretation as deviation of a relative risk from its reference value

$$\beta_D^{\langle\gamma\rangle}(u) = \log \frac{RR_u(Y^D = 1)}{\mathcal{RR}_u(Y^D = 1)}. \tag{20}$$

We close this section by noticing that also the LM and the LML regression models belong to the family of generalized log-linear models. This allows us to exploit

the asymptotic results of Lang (1996) for the computation of maximum likelihood estimates (MLEs) and for model comparison. Nevertheless, we remark that the regression framework introduced in this paper is entirely novel. Both the LM and the LML link functions allow us to specify the functional relationship between $\pi$ and the regression coefficients in closed form. This is a specific property that is not shared by the wider class of generalised log-linear models, and that makes it possible to specify closed-form functional relationships between regression coefficients of different responses. In turn, this is the basis for the factorization of relative risks with respect to the corresponding relative risks of lower order regressions.

# 5 Analysis of multimorbidity data

We now apply the LM and the LML regression models to the analysis of the multimorbidity data described in Section 2. We consider four response variables: $Y_b$ = *Bone fracture*, $Y_c$ = *Cardiovascular disease*, $Y_d$ = *Diabetes* and $Y_r$ = *Renal failure*. Hence, $V = \{b, c, d, r\}$ and $Y_V$ takes values in $\{0, 1\}^4$ where, for each variable, the value 1 encodes the presence of the disease. The four responses define 11 multimorbidity patterns denoted by the subsets $D \subseteq V$ with $|D| \geq 2$ and we say that $k = |D|$ is the *size* of the multimorbidity pattern.

For the computation of MLEs we applied the algorithm and the maximization procedure given in Lang (1996) properly adjusted to account for the inclusion of the LM and LML link functions; for technical details and a review of further maximization approaches see also Evans and Forcina (2013) and references therein.

It is well established that asymptotic methods are not efficient when the table of observed counts is sparse, that is when many cells have small frequencies (see Agresti, 2013, Section 10.6). For a given sample size, sparsity increases with the number of variables included in the analysis, therefore inference is less reliable for response associatons of higher-order since they are computed on the relevant marginal tables. In order to keep sparsity at an acceptable level, we restrict the analysis to the effect of two binary covariates indexed by $U = \{a, h\}$; specifially, $X_h = HIV$ with the level 1 encoding the presence of the infection and $X_a = Age$ with the value 1 for patients aged 45 or more. Firstly, in Subsection 5.1, we consider the regression model including only the covariate $X_h$, so that each regression coefficient represents the marginal effect of HIV on a $D$-response associations. Next, in Subsection 5.2 we consider a regression model with two covariates including also the effect of age.

We remark that the case-control design for this study is not based on an outcome-

dependent sampling. In fact the enrollment of each patient is independent of the disease status given the HIV status and further individual covariates. Therefore, for this case-study the population relative risk represents an identifiable measure of association, unlike case-control designs where the sample selection depends on the outcome of interest.

## 5.1 The single covariate case

When only the covariate $X_h$ is included in the model, the regression coefficients have a straightforward interpretation in terms of relative risks for cases versus controls. Indeed, in the LM regression model, it holds that for $|D| = 1$ the coefficient $\beta_D^{\langle\mu\rangle}(h)$ is the log-relative risk of the occurrence of single commorbidities and, otherwise, it is the log-relative risk of the occurrence of the multimorbidity pattern $D$; see (8). In the LML regression, the coefficient $\beta_D^{\langle\gamma\rangle}(h)$ is equal to $\beta_D^{\langle\mu\rangle}(h)$ in case of single responses, otherwise it is the log-ratio of the relative and the reference relative risks for the occurrence of the multimorbidity pattern $D$, as shown in (20). Hence, from Proposition 4.8, the no-effect of HIV defined by $\beta_D^{\langle\gamma\rangle}(h) = 0$ with $|D| > 1$ implies that the relative risk equals the reference relative risk for the multimorbidity pattern $D$. Conversely, a positive or a negative value of $\beta_D^{\langle\gamma\rangle}(h)$ states that the relative risk for the pattern $D$ is higher or lower than its reference relative risk, respectively.

As a preliminary analysis, we provide in Table 1 the MLEs under the saturated LM and LML regression models. To clarify the meaning of the values given in this table consider, for instance, the disease pattern $\{b, c\}$, that is *bone fracture–diabetes*. The estimated LM coefficient $\hat{\beta}_{bc}^{\langle\mu\rangle}(h) = 2.941$ provides a strong evidence of the existence of an effect of HIV on the co-occurrence of the two diseases whereas the corresponding LML parameter estimate $\hat{\beta}_{bc}^{\langle\gamma\rangle}(h) = -0.737$ is not significantly different from zero, and this can be interpreted as no-effect of HIV on the association of the two diseases. More specifically, the estimated relative risk of the co-occurrence of bone fracture and diabetes takes value $\exp\{\hat{\beta}_{bc}^{\langle\mu\rangle}(h)\} = \exp(2.941) = 18.9$ and is significantly different from 1; however, it is not significantly different from the product of the two marginal relative risks given by $\exp\{\beta_b^{\langle\mu\rangle}(h)\} \times \exp\{\beta_c^{\langle\mu\rangle}(h)\}$. The latter is the value taken by the relative risk of the co-occurrence of the two diseases when $Y_b \perp\!\!\!\perp Y_c | X_h$ and therefore it implies that the relative risk of the co-occurrence of the two diseases only depends on the marginal relative risks of the two diseases rather than on the way the two diseases associate to give the multimorbidity pattern. Different is the case of the disease pattern $\{c, d\}$ because in this case the estimated LML coefficient

$\hat{\beta}_{cd}^{\langle\gamma\rangle}(h) = -1.165$ is significantly different form zero and negative and therefore the relative risk of the co-occurrence of the two diseases is significantly smaller than expected in the case the two diseases are conditionally independent given HIV. We remark that, when we say that a regression coefficient is significantly different from zero we refer to the statistical test, at level 5%, based on the asymptotic normal distribution of the MLEs.

As expected (see the discussion following Proposition 4.1) in Table 1 the estimated log-relative risks $\hat{\beta}_D^{\langle\mu\rangle}(h)$ are positive for every single disease and for every multimorbidity pattern. Furthermore, all the corresponding regression coefficients are significantly different from zero, except for the pattern $\{b, c, d\}$. The lowest relative risk corresponds to the pattern $\{c, d\}$, whereas the highest relative risks are in correspondence of the patterns $\{b, r\}$, $\{c, r\}$ and $\{b, d, r\}$. More informative is the analysis of the LML regression coefficients. Indeed, most of the coefficients $\beta_D^{\langle\gamma\rangle}(h)$ are not significantly different from zero, thereby providing an empirical evidence that the log-relative risks for the corresponding disease patterns are positive as a consequence of the effect of HIV on the single diseases they are composed, rather than for an effect of HIV on the associations of diseases. It is interesting that most of the coefficients $\hat{\beta}_D^{\langle\gamma\rangle}(h)$, for $|D| \geq 2$, take a negative value. This is especially true for the multimorbidity patterns of size two, suggesting that the corresponding relative risks, although positive, take a value that is smaller than the value one would expect in case of conditional independence of the diseases.

The results provided by the saturated model can be summarized considering the average HIV-effect for disease patterns of the same size $k$, under the LM and LML approaches. In particular, we compute the average of $\hat{\beta}_D^{\langle\mu\rangle}(h)$ and of $\hat{\beta}_D^{\langle\gamma\rangle}(h)$, for $D$-response associations of the same size; see Appendix B for further details. Confidence intervals for these effects are plotted in Figure 1. The plot in the left panel of Figure 1 gives the estimated LM average effect, i.e. the average log-relative risk for the occurrence of multimorbidity patterns of the same size, which clearly increases with the size of the disease pattern. On the other hand, the plot in the right panel of Figure 1 shows the estimated LML average effect, that is the average effect of HIV on the association among diseases forming patterns of the same size. This effect appears to be negative for patterns of size $k = 2$ and it might be null for $k = 3$ and $k = 4$.

Next, we apply the forward inclusion stepwise procedure described in Appendix B and select, in this way, the LML regression submodel given in Table 2. The latter provides a very good fit of the data with a deviance 3.45 on 12 degrees of freedom and $p$-value, computed on the asymptotic chi-square distribution of the deviance,

Table 1: The saturated LML and LM regression models for $Y_V|X_h$. The table gives the MLEs of the regression coefficients with their standard error (s.e.) and $p$-value.

| $D$ | $\hat{\beta}_D^{\langle\gamma\rangle}(\emptyset)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\mu\rangle}(\emptyset)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\mu\rangle}(h)$ | s.e. | $p$-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\{b\}$ | -4.573 | 0.106 | $<.001$ | 2.621 | 0.115 | $<.001$ | -4.573 | 0.106 | $<.001$ | 2.621 | 0.115 | $<.001$ |
| $\{c\}$ | -4.476 | 0.101 | $<.001$ | 1.056 | 0.143 | $<.001$ | -4.476 | 0.101 | $<.001$ | 1.056 | 0.143 | $<.001$ |
| $\{d\}$ | -3.255 | 0.054 | $<.001$ | 1.061 | 0.075 | $<.001$ | -3.255 | 0.054 | $<.001$ | 1.061 | 0.075 | $<.001$ |
| $\{r\}$ | -6.321 | 0.251 | $<.001$ | 3.570 | 0.261 | $<.001$ | -6.321 | 0.251 | $<.001$ | 3.570 | 0.261 | $<.001$ |
| $\{b,c\}$ | 1.158 | 0.524 | 0.027 | -0.737 | 0.558 | 0.187 | -7.892 | 0.541 | $<.001$ | 2.941 | 0.584 | $<.001$ |
| $\{b,d\}$ | 0.422 | 0.415 | 0.309 | -0.539 | 0.437 | 0.217 | -7.407 | 0.430 | $<.001$ | 3.144 | 0.457 | $<.001$ |
| $\{b,r\}$ | 0.230 | 0.272 | 0.398 | -0.257 | 0.325 | 0.429 | -10.665 | 0.005 | $<.001$ | 5.935 | 0.198 | $<.001$ |
| $\{c,d\}$ | 1.776 | 0.181 | $<.001$ | -1.165 | 0.268 | $<.001$ | -5.955 | 0.211 | $<.001$ | 0.953 | 0.309 | 0.002 |
| $\{c,r\}$ | 0.133 | 0.270 | 0.622 | 0.404 | 0.393 | 0.304 | -10.665 | 0.005 | $<.001$ | 5.030 | 0.310 | $<.001$ |
| $\{d,r\}$ | 1.684 | 0.483 | $<.001$ | -0.863 | 0.497 | 0.082 | -7.892 | 0.541 | $<.001$ | 3.768 | 0.560 | $<.001$ |
| $\{b,c,d\}$ | -0.011 | 0.697 | 0.987 | -0.553 | 0.909 | 0.543 | -8.960 | 0.909 | $<.001$ | 1.745 | 1.131 | 0.123 |
| $\{b,c,r\}$ | 2.493 | 0.581 | $<.001$ | -1.846 | 0.676 | 0.006 | -11.358 | 0.005 | $<.001$ | 4.812 | 0.487 | $<.001$ |
| $\{b,d,r\}$ | 0.456 | 0.634 | 0.472 | -0.101 | 0.681 | 0.882 | -11.358 | 0.005 | $<.001$ | 5.493 | 0.349 | $<.001$ |
| $\{c,d,r\}$ | -0.898 | 0.513 | 0.080 | 0.748 | 0.616 | 0.225 | -11.358 | 0.005 | $<.001$ | 4.812 | 0.487 | $<.001$ |
| $\{b,c,d,r\}$ | -0.867 | 0.846 | 0.305 | 0.742 | 1.015 | 0.465 | -12.051 | 0.005 | $<.001$ | 4.143 | 0.952 | $<.001$ |

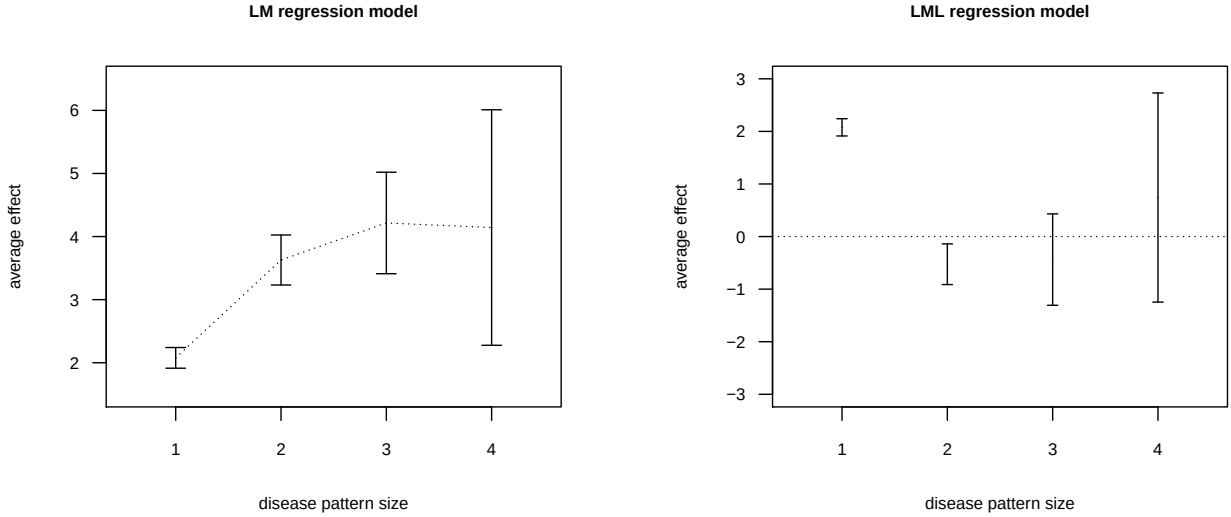


Figure 1: Confidence intervals of the average HIV-effect for response associations for disease patterns of the same size under the saturated LM and LML regression models.

Table 2: The selected LML, and the corresponding LM, regression model for $Y_V|X_h$. The table gives the MLEs of the regression coefficients with their standard error (s.e.) and $p$-value.

| $D$ | $\hat{\beta}_D^{\langle\gamma\rangle}(\emptyset)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\mu\rangle}(\emptyset)$ | s.e. | $\hat{\beta}_D^{\langle\mu\rangle}(h)$ | s.e. |
|---|---|---|---|---|---|---|---|---|---|---|
| $\{b\}$ | -4.534 | 0.088 | $<.001$ | 2.579 | 0.100 | $<.001$ | -4.534 | 0.088 | 2.579 | 0.100 |
| $\{c\}$ | -4.474 | 0.093 | $<.001$ | 1.056 | 0.138 | $<.001$ | -4.474 | 0.093 | 1.056 | 0.138 |
| $\{d\}$ | -3.253 | 0.053 | $<.001$ | 1.058 | 0.075 | $<.001$ | -3.253 | 0.053 | 1.058 | 0.075 |
| $\{r\}$ | -6.131 | 0.089 | $<.001$ | 3.383 | 0.114 | $<.001$ | -6.131 | 0.089 | 3.383 | 0.114 |
| $\{b,c\}$ | 0.443 | 0.181 | 0.014 | · | · | · | -8.565 | 0.227 | 3.635 | 0.178 |
| $\{b,d\}$ | · | · | · | · | · | · | -7.787 | 0.101 | 3.637 | 0.122 |
| $\{b,r\}$ | · | · | · | · | · | · | -10.665 | 0.011 | 5.962 | 0.086 |
| $\{c,d\}$ | 1.777 | 0.124 | $<.001$ | -1.232 | 0.235 | $<.001$ | -5.950 | 0.182 | 0.882 | 0.291 |
| $\{c,r\}$ | · | · | · | · | · | · | -10.605 | 0.118 | 4.439 | 0.172 |
| $\{d,r\}$ | 0.840 | 0.102 | $<.001$ | · | · | · | -8.544 | 0.155 | 4.440 | 0.142 |
| $\{b,c,d\}$ | · | · | · | · | · | · | -10.041 | 0.261 | 3.461 | 0.299 |
| $\{b,c,r\}$ | 3.337 | 0.202 | $<.001$ | -2.631 | 0.431 | $<.001$ | -11.358 | 0.006 | 4.387 | 0.439 |
| $\{b,d,r\}$ | 1.720 | 0.112 | $<.001$ | -1.382 | 0.238 | $<.001$ | -11.358 | 0.011 | 5.638 | 0.251 |
| $\{c,d,r\}$ | · | · | · | · | · | · | -11.241 | 0.222 | 4.264 | 0.323 |
| $\{b,c,d,r\}$ | -1.777 | 0.124 | $<.001$ | 1.284 | 0.530 | 0.015 | -12.051 | 0.006 | 4.115 | 0.729 |

equal to 0.99. In the selected model there is no effect of HIV on the associations $\{b,c\}$, $\{b,d\}$, $\{b,r\}$, $\{c,r\}$, $\{d,r\}$, $\{b,c,d\}$ and $\{c,d,r\}$. Furthermore, it follows from Theorem 4.2 and Proposition 4.8 that, under the selected model, three pairs of responses are conditionally independent given HIV, specifically, $Y_b \perp\!\!\!\perp Y_d|X_h$, $Y_b \perp\!\!\!\perp Y_r|X_h$ and $Y_c \perp\!\!\!\perp Y_r|X_h$.

We now look more closely at the values taken by the estimated regression coefficients given in Table 2. We base this analysis on the asymptotic normal distribution of the MLEs, but it is important to remark that, as we are dealing with post-selection parameter estimates, there might be distortions on the sampling distributions; see Berk et al. (2013) for a discussion and recent developments. The highest estimated relative risks are for multimorbidity patterns $\{b,r\}$ and $\{b,d,r\}$ with 95% confidence intervals $(5.793; 6.130)$ and $(5.146; 6.130)$, respectively. More generally, high relative risks are given in correspondence of multimorbidity patterns including *Renal failure*.

The negative values taken by the estimates $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ for the remaining patterns of size two and three suggest that the relative risks of the multimorbidity patterns $\{c,d\}$, $\{b,d,r\}$ and $\{b,c,r\}$ are lower than their reference relative risks. For the disease pattern of size 4 the estimate of $\beta_D^{\langle\gamma\rangle}(h)$ is positive suggesting that the relative risk of the pattern $\{b,c,d,r\}$ is higher than its reference relative risk.

## 5.2 The two-covariate case

We now introduce in the analysis of the previous section the additional covariate $X_a = Age$ and apply the model selection procedure described in Appendix B to obtain the model given in Table 3. Such model has deviance 32.996 on 33 degrees of freedom ($p = 0.467$).

Table 3: The selected LML, and the corresponding LM, regression model for $Y_V|X_{ah}$. The table gives the MLEs of the regression coefficients with their standard error (s.e.) and $p$-value.

| $D$ | $\hat{\beta}_D^{\langle\gamma\rangle}(\emptyset)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\gamma\rangle}(a)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\mu\rangle}(h)$ | s.e. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\{b\}$ | -4.717 | 0.117 | $<.001$ | 0.274 | 0.087 | 0.002 | 2.604 | 0.113 | $<.001$ | 2.604 | 0.113 |
| $\{c\}$ | -5.347 | 0.170 | $<.001$ | 1.268 | 0.175 | $<.001$ | 1.044 | 0.140 | $<.001$ | 1.044 | 0.140 |
| $\{d\}$ | -4.052 | 0.093 | $<.001$ | 1.159 | 0.095 | $<.001$ | 1.059 | 0.074 | $<.001$ | 1.059 | 0.074 |
| $\{r\}$ | -6.745 | 0.223 | $<.001$ | 0.869 | 0.153 | $<.001$ | 3.419 | 0.219 | $<.001$ | 3.419 | 0.219 |
| $\{b,c\}$ | 1.292 | 0.273 | $<.001$ | · | · | · | -0.907 | 0.329 | 0.006 | 2.742 | 0.363 |
| $\{b,d\}$ | · | · | · | · | · | · | · | · | · | 3.663 | 0.135 |
| $\{b,r\}$ | 2.732 | 0.287 | $<.001$ | -0.720 | 0.322 | 0.025 | -2.223 | 0.361 | $<.001$ | 3.799 | 0.375 |
| $\{c,d\}$ | 2.421 | 0.308 | $<.001$ | -0.980 | 0.312 | 0.002 | -1.119 | 0.257 | $<.001$ | 0.985 | 0.297 |
| $\{c,r\}$ | 2.252 | 0.262 | $<.001$ | · | · | · | -1.881 | 0.379 | $<.001$ | 2.582 | 0.348 |
| $\{d,r\}$ | 2.249 | 0.294 | $<.001$ | -0.628 | 0.279 | 0.024 | -1.053 | 0.320 | 0.001 | 3.425 | 0.348 |
| $\{b,c,d\}$ | · | · | · | · | · | · | · | · | · | 2.682 | 0.442 |
| $\{b,c,r\}$ | · | · | · | · | · | · | · | · | · | 2.056 | 0.474 |
| $\{b,d,r\}$ | · | · | · | · | · | · | · | · | · | 3.805 | 0.389 |
| $\{c,d,r\}$ | -1.312 | 0.376 | $<.001$ | · | · | · | 1.299 | 0.485 | 0.007 | 2.769 | 0.579 |
| $\{b,c,d,r\}$ | · | · | · | · | · | · | · | · | · | 2.242 | 0.661 |

When the model includes more than one covariate, the regression coefficients can be used to compute conditional relative risks as shown in Lemma 4.3 and Corollary 4.9. However, in the model resulting from the application of the selection procedure the interactions $\beta_D^{\langle\gamma\rangle}(ah)$ are equal to zero for every $D \subseteq V$ and, by (1) and (17), this implies that also $\beta_D^{\langle\mu\rangle}(ah)$ are equal to zero for every $D \subseteq V$. As a consequence, in the selected model the regression coefficients have a direct interpretation in term of conditional relative risks, for every $D \subseteq V$, as follows

$$\beta_D^{\langle\mu\rangle}(h) = \log RR_h(Y^D = 1|\emptyset) = \log RR_h(Y^D = 1|a)$$

and similarly for $\beta_D^{\langle\mu\rangle}(a)$. LML regression coefficients coincide with LM regression

coefficients for $|D| \leq 1$ whereas, for $|D| > 1$, if follows from (4.9) that

$$\beta_D^{\langle\gamma\rangle}(h) = \log \frac{RR_h(Y^D = 1|\emptyset)}{\mathcal{RR}_h(Y^D = 1|\emptyset)} = \log \frac{RR_h(Y^D = 1|a)}{\mathcal{RR}_h(Y^D = 1|a)}$$

and similarly for $\beta_D^{\langle\gamma\rangle}(a)$.

Table 3 includes, for every $D \subseteq V$, the estimates $\hat{\beta}_D^{\langle\mu\rangle}(h)$ of the corresponding LM regression model. As well as in the model with one covariate, also in this case the fitted model provides high relative risks of patterns involving *Renal failure*. Nevertheless, compared with the results illustrated in Section 5.1, the inclusion of *Age* leads to a sensitive reduction of the estimated relative risk $\hat{\beta}_D^{\langle\mu\rangle}(h)$ for most of the multimorbidity patterns.

The fitted model shows a negative value of the estimates $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ for most of the patterns of size two, except for the pattern $D = \{b, d\}$ where the model constraints imply the conditional independence relationship $Y_b \perp\!\!\!\perp Y_d | X_{ah}$. According to the selected model, HIV has no-effect on every $D$-response association of size $|D| > 2$ with the exception of the pattern $\{c, d, r\}$.

The estimates of the LML regression coefficients $\hat{\beta}_D^{\langle\gamma\rangle}(a)$ show a positive value in univariate regressions as they coincide with the estimates of the log-relative risk of a single comorbidity for the increasing of age; in particular, the highest estimates are for the events $\{Y_c = 1\}$ and $\{Y_d = 1\}$, whereas the lowest is for $\{Y_b = 1\}$. For most of the multimorbidity patterns, the dataset supports the hypotheses of no-effect of age with $\hat{\beta}_D^{\langle\gamma\rangle}(a) = 0$, except for patterns $\{b, r\}$, $\{c, d\}$ and $\{d, r\}$ where the estimates of the corresponding coefficients are negative.

# 6 Discussion

A wide range of alternative link functions for binary data have been proposed in the literature and, in particular, the marginal log-linear approach of Bergsma and Rudas (2002) provides a wide and flexible class of links including the multivariate logistic one. However, in marginal regression modeling, iterative procedures need to be typically used to compute the cell probabilities from the parameters of the model. A key property of both the LM and the LML link functions is that probabilities can be analytically computed from the parameters. This inverse closed-form mapping is a distinguishing feature that confers our approach fundamental advantages that are not generally shared by other methods. In particular, this makes it possible to

derive a closed-form functional relationships between the coefficients of regressions of different order that turns out to be very convenient when the interest is on the association of responses as in the multimorbidity application or, more generally, as stated by the line of enquiry (iii) given in McCullagh and Nelder (1989, Sec. 6.5); see also Section 1. More concretely, suitable collections of zero LML coefficients imply the analysis in terms of relative risks can be equivalently performed in marginal distributions of the responses, even when independence relationships between responses do not hold true. Similar conclusions cannot be drawn by looking, for instance, at the coefficients of regressions based on the multivariate logistic link where analysis in terms of covariate effect on response associations can be equivalently carried out in marginal distributions only in case of independence.

With respect to the lines of enquiry given in Section 1, we remark that although the main focus of this paper is on the line of enquiry (iii), the LM and the LML regression models are useful also when the interest is for the lines of enquiry (i) and (ii). Firstly, with respect to (i), that is the analysis of the dependence structure of each response marginally on covariates, a central role is played by multivariate logistic regression because it maintains a marginal logistic regression interpretation for the single outcomes. However, when the interest is for relative risks, rather than odds ratios, the LM and LML regression models provide a useful alternative. Secondly, when (ii) is of concern, that is a model for the joint distribution of all responses, Proposition 4.1 and Corollary 4.7 show that the LM and LML regression models allow one to identify independencies among subsets of responses, conditionally on covariates. Interestingly, this feature is shared by the multivariate logistic regression (Marchetti and Lupparelli, 2011) and, more precisely, the family of regression graph models (see Wermuth and Sadeghi, 2012) for binary data turns out as a special case of our approach. However, the regression graph representation of the model does not give the associations of responses and therefore we do not pursue this aspect here.

Future research directions for the class of LM and LML regression models involve the extension to response variables with an arbitrary number of levels, following a similar generalization of the LM and the LML parameterizations provided by Roverato (2015). It is also of interest the inclusion of continuous covariates and of additive random effects which may be useful in the case where the sampling design induces unobserved correlation between units which cannot be totally explained by the covariates. Unlike other approaches to regressions for categorical responses, such as the GEE models, the LM and LML regressions do not seem to lend themselves to semi-parametric fitting approaches because correlations between responses are regarded as

parameters of interest rather than as nuisance parameters.


# Acknowledgments

We gratefully acknowledge Mauro Gasparini, Luca La Rocca and Nanny Wermuth for helpful discussions. We thank two Referees for their helpful comments.


# A Proofs

## Proof of Lemma 3.1

(i) $\Rightarrow$ (ii). Since $\theta_D(E) = \sum_{E' \subseteq E} \beta_D^{\langle\theta\rangle}(E')$, then if $E_1, E_2 \subseteq U$ it follows by (i) both that

$$\theta_D(E_1) = \sum_{E' \subseteq E_1} \beta_D^{\langle\theta\rangle}(E') = \sum_{E' \subseteq E_1 \setminus U'} \beta_D^{\langle\theta\rangle}(E')$$

and that

$$\theta_D(E_2) = \sum_{E' \subseteq E_2} \beta_D^{\langle\theta\rangle}(E') = \sum_{E' \subseteq E_2 \setminus U'} \beta_D^{\langle\theta\rangle}(E').$$

Hence, $\theta_D(E_1) = \theta_D(E_2)$ whenever $E_1 \setminus U' = E_2 \setminus U'$, as required.

We now show that (ii) $\Rightarrow$ (i). If $E \subseteq U$ is such that $E \cap U' \neq \emptyset$ then we can find an element $u \in E \cap U'$ and, furthermore, it is straightforward to see that for every $E' \subseteq E \setminus \{u\}$ it holds that $E' \setminus U' = (E' \cup \{u\}) \setminus U'$ and this implies, by (ii), that $\theta_D(E') = \theta_D(E' \cup \{u\})$. The result follows because

$$\begin{aligned}
\beta_D^{\langle\theta\rangle}(E) &= \sum_{E' \subseteq E} (-1)^{|E \setminus E'|} \theta_D(E') \\
&= \sum_{E' \subseteq E \setminus \{v\}} (-1)^{|E \setminus E'|} \left\{\theta_D(E') - \theta_D(E' \cup \{v\})\right\} \\
&= 0.
\end{aligned}$$

## Proof of Proposition 4.1

Recall that, since by construction $\beta_\emptyset^{\langle\mu\rangle}(E) = 0$ for every $E \subseteq U$ it is sufficient to consider the case where $D' \neq \emptyset$. We first show that $\beta_{D'}^{\langle\mu\rangle}(E) = 0$ for every $\emptyset \neq D' \subseteq D$

and $E \subseteq U$ such that $E \cap U' \neq \emptyset$ implies that the conditional distribution

$$\mathrm{pr}(Y_H = 1, Y_{D\setminus H} = 0 \mid X_E = 1, X_{U\setminus E} = 0) \tag{21}$$

does not depend on the value taken by $X_{U'}$ for every $H \subseteq D$ and $E \subseteq U$. The result follows by noticing that for every $H \subseteq D$ and $E \subseteq U$ it holds that

$$\mathrm{pr}(Y_H = 1, Y_{D\setminus H} = 0 \mid X_E = 1, X_{U\setminus E} = 0) = \sum_{H' \subseteq D\setminus H} (-1)^{|H'|}\, \mu_{H\cup H'}(E) \tag{22}$$

and that every term $\mu_{H\cup H'}(E)$ in (22) does not depend on the value taken by $X_{U'}$. This follows from Lemma 3.1 which states that $\beta^{\langle\mu\rangle}_{D'}(E) = 0$, for every $\emptyset \neq D' \subseteq D$ and $E \subseteq U$ with $E \cap U' \neq \emptyset$, implies $\mu_{D'}(E_1) = \mu_{D'}(E_2)$, for every $E_1, E_2 \subseteq U$ with $E_1\setminus U' = E_2\setminus U'$. In turns this means that every term $\mu_{H\cup H'}(E)$ in (22) with $H \cup H' \neq \emptyset$ does not depend on the value taken by $X_{U'}$. Furthermore, if $H \cup H' = \emptyset$, then $\mu_{H\cup H'}(E) = 1$ by definition and this completes the first part of the proof.

We now prove the reverse implication, that is we assume that the conditional distribution of $Y^D|X_U$ in (21) does not depend on $X_{U'}$ and show that this implies $\beta^{\langle\mu\rangle}_{D'}(E) = 0$ for every $\emptyset \neq D' \subseteq D$ and $E \subseteq U$ such that $E \cap U' \neq \emptyset$. If $\emptyset \neq D' \subseteq D$ then, by assumption, the conditional distribution of $Y^{D'}|X_U$ does not depend on $X_{U'}$ and, in turn, this implies that $\mu_{D'}(E_1) = \mu_{D'}(E_2)$ for every $E_1, E_2 \subseteq U$ such that $E_1\setminus U' = E_2\setminus U'$ because, in this case, $\mu_{D'}(E_1)$ and $\mu_{D'}(E_2)$ are the probabilities of $Y_{D'} = 1$ conditioned on values of $X_U$ that may only differ for variables in $X_{U'}$. Hence we can apply Lemma 3.1 to obtain that $\beta^{\langle\mu\rangle}_{D'}(E) = 0$ for every $E \subseteq U$ such that $E \cap U' \neq \emptyset$.

## Proof of Theorem 4.2

Theorem 1 of Roverato et al. (2013) implies that $Y_A \perp\!\!\!\perp Y_B | X_U$ if and only if every row of $M^\top \log(\mu)$ indexed by $D \subseteq A \cap B$ with $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ is equal to zero. Since $\log(\mu) = \beta^{\langle\mu\rangle} Z_U$ then the row of $M^\top \log(\mu)$ indexed by $D$ is equal to zero if and only if the corresponding row of $M^\top \beta^{\langle\mu\rangle} Z_U$ is equal to zero. In turn, a row of $M^\top \beta^{\langle\mu\rangle} Z_U$ is equal to zero if and only if the corresponding row of $M^\top \beta^{\langle\mu\rangle}$ is equal to zero, because $Z_U$ has full rank. The row of $M^\top \beta^{\langle\mu\rangle}$ indexed by $D$ has entries $\sum_{D' \subseteq D} (-1)^{|D\setminus D'|}\, \beta^{\langle\mu\rangle}_{D'}(E) = \beta^{\langle\mu\rangle}_{D}(E) - \mathcal{B}^{\langle\mu\rangle}_{D}(E)$ for $E \subseteq U$ and, therefore, it is equal to zero if and only if $\beta^{\langle\mu\rangle}_{D}(E) = \mathcal{B}^{\langle\mu\rangle}_{D}(E)$ for every $E \subseteq U$, as required.

## Proof of Lemma 4.3

The first equality follows immediately from the fact that the relative risk is the ratio of two mean parameters as follows

$$\log RR_u(Y^D = 1|E) = \log \left\{ \frac{\mu_D(E \cup \{u\})}{\mu_D(E)} \right\} = \sum_{E' \subseteq E} \beta_D^{\langle \mu \rangle}(E' \cup \{u\}).$$

The second equality can be proved as follows,

$$\begin{aligned}
\sum_{E' \subseteq E} \mathcal{B}_D^{\langle \mu \rangle}(E' \cup \{u\}) &= -\sum_{E' \subseteq E} \sum_{D' \subset D} (-1)^{|D \setminus D'|} \, \beta_{D'}^{\langle \mu \rangle}(E' \cup \{u\}) \\
&= -\sum_{D' \subset D} (-1)^{|D \setminus D'|} \sum_{E' \subseteq E} \beta_{D'}^{\langle \mu \rangle}(E' \cup \{u\}) \\
&= -\sum_{D' \subset D} (-1)^{|D \setminus D'|} \log RR_u(Y^{D'} = 1|E) \\
&= \log \prod_{D' \subset D} RR_u(Y^{D'} = 1|E)^{(-1)^{|D \setminus D'|+1}} \\
&= \log \mathcal{RR}_u(Y^D = 1|E).
\end{aligned}$$

## Proof of Corollary 4.4

By Lemma 4.3

$$\log RR_u(Y^D = 1|E) = \sum_{E' \subseteq E} \beta_D^{\langle \mu \rangle}(E' \cup \{u\})$$

and it follows from Theorem 4.2, the fact that $Y_A \perp\!\!\!\perp Y_B | X_U$, and Lemma 4.3 that

$$\log RR_u(Y^D = 1|E) = \sum_{E' \subseteq E} \mathcal{B}_D^{\langle \mu \rangle}(E' \cup \{u\}) = \log \mathcal{RR}_u(Y^D = 1|E).$$

## Proof of Corollary 4.5

The fact that (13)⇒(15) follows immediately form Lemma 4.3. To show the reverse implication, that is (15)⇒(13) we notice that in this case Lemma 4.3 implies

$$\sum_{E' \subseteq E} \beta_D^{\langle \mu \rangle}(E' \cup \{u\}) = \sum_{E' \subseteq E} \mathcal{B}_D^{\langle \mu \rangle}(E' \cup \{u\}) \text{ for every } E \subseteq U \setminus \{u\} \tag{23}$$

Hence, for $E = \emptyset$, (23) implies $\beta_D^{\langle\mu\rangle}(\{u\}) = \mathcal{B}_D^{\langle\mu\rangle}(\{u\})$, and the proof follows by induction on the dimension of $E$. We consider $E \subseteq U\backslash\{u\}$, with $E \neq \emptyset$, and show that if the result is true for every $E' \subset E$ then it also holds true for $E$. Indeed, in this case, it follows from (23) that

$$\sum_{E' \subset E} \mathcal{B}_D^{\langle\mu\rangle}(E' \cup \{u\}) + \beta_D^{\langle\mu\rangle}(E \cup \{u\}) = \sum_{E' \subseteq E} \mathcal{B}_D^{\langle\mu\rangle}(E' \cup \{u\})$$

so that $\beta_D^{\langle\mu\rangle}(E \cup \{u\}) = \mathcal{B}_D^{\langle\mu\rangle}(E \cup \{u\})$.

## Proof of Lemma 4.6

By definition, $\gamma = M_V^\top \log(\mu)$ so that $\gamma M_U = M_V^\top \log(\mu) M_U$ and the result follows by recalling that $\beta^{\langle\gamma\rangle} = \gamma M_U$ and $\beta^{\langle\mu\rangle} = \log(\mu) M_U$.

## Proof of Corollary 4.7

This is an immediate consequence of Proposition 4.1 and Lemma 4.6 because if, for $E \subseteq U$ it holds that $\beta_{D'}^{\langle\mu\rangle}(E) = 0$ for every $D' \subseteq D$, then it follows from (17) that also $\beta_{D'}^{\langle\gamma\rangle}(E) = 0$ for every $D' \subseteq D$. Conversely, it follows from Lemma 4.6 that $\beta_D^{\langle\mu\rangle}(E) = \sum_{D' \subseteq D} \beta_{D'}^{\langle\gamma\rangle}(E)$ and, therefore, if for $E \subseteq U$ it holds that $\beta_{D'}^{\langle\gamma\rangle}(E) = 0$ for every $D' \subseteq D$ then also $\beta_{D'}^{\langle\mu\rangle}(E) = 0$ for every $D' \subseteq D$.

## Proof of Corollary 4.9

This follows from Proposition 4.8 that

$$\sum_{E' \subseteq E} \beta_D^{\langle\gamma\rangle}(E) = \sum_{E' \subseteq E} \beta_D^{\langle\mu\rangle}(E) - \sum_{E' \subseteq E} \mathcal{B}_D^{\langle\mu\rangle}(E)$$

and the result follows from Lemma 4.3.

Table 4: Average HIV-effect on $D$-response asssociations of the same size $k = 1, \ldots, 4$, under the saturated LML and LM regression model for $Y_V|X_h$. Standard errors are included in brackets.

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\bar{\beta}_k^{\langle\gamma\rangle}(h)$ | 1.888(0.064) | −0.693(0.203) | −0.361(0.460) | 0.742(1.015) |
| $\bar{\beta}_k^{\langle\mu\rangle}(h)$ | 1.888(0.064) | 3.272(0.225) | 4.614(0.337) | 4.143(0.952) |

# B Technical details on the analysis of the multimorbidity data

## The single covariate case

The confidence intervals in Figure 1 for the average effects of HIV on $D$-response associations of same size $k = |D|$ are obtained from the weighted averages of the MLEs of the saturated LM and LML regression models given in Table 1. For every $k = 1, 2, 3, 4$ we set

$$\bar{\beta}_k^{\langle\mu\rangle}(h) = \sum_{D\subseteq V:|D|=k} w_D \hat{\beta}_D^{\langle\mu\rangle}(h) \quad \text{and} \quad \bar{\beta}_k^{\langle\gamma\rangle} = \sum_{D\subseteq V:|D|=k} w_D \hat{\beta}_D^{\langle\gamma\rangle}(h),$$

where $w_D$, for $\emptyset \neq D \subseteq V$ is a suitable collection of weights. Concretely, given the observed sample, we set $w_D$ equal to the number of patients affected by disease pattern $D$ normalized within the group of patients affected by disease patterns of the same size. Hence, each $w_D$ represents the observed proportion of patients affected by the multimorbidity event $Y^D$ within the subgroup of patients affected by multimorbidity events of the same size $|D|$ so that, for every $k$, $\sum_{D\subseteq V:|D|=k} w_D = 1$. It is worth noticing that $\bar{\beta}_1^{\langle\mu\rangle}(h) = \bar{\beta}_1^{\langle\gamma\rangle}(h)$ because for univariate regressions with $k = 1$ the LM and the LML coefficients are equivalent, and $\bar{\beta}_4^{\langle\mu\rangle}(h) = \hat{\beta}_V^{\langle\mu\rangle}(h)$ and $\bar{\beta}_4^{\langle\gamma\rangle}(h) = \hat{\beta}_V^{\langle\gamma\rangle}(h)$ as $V$ is the only pattern of size four. Table 4 collects these average effects with their standard errors and, to better understand the role played by these weights, one can check from Table 1 that the weighted means are closer to the medians of the averaged regression coefficients rather than to their arithmetic means.

The procedure adopted for model selection exploits the upward compatibility property satisfied by the LML parameterization which implies that every term $\gamma_D(\cdot)$ of $Y_V|X_h$ can be computed in the relevant marginal distribution of $Y_D|X_h$, with $D \subseteq V$. Then, starting from univariate LML regression models, a forward procedure is fol-

lowed which updates step-by-step the regression model for response associations of higher order such that models for associations of lower order (selected in the relevant marginal distributions) are preserved. In details, the procedure is based on four ordered steps: at every step $i$, with $i = 1, \ldots, 4$, a LML regression models for $Y_D|X_h$ is selected for every $D \subseteq V$ with $|D| = i$. At every step, the structure, that is the zero terms, of the models chosen at the previous steps is maintained. Reduced models at each step are selected setting zero constraints on regression coefficients not significatively different from zero. Table 5, 6 and 7 give the MLEs of the LML regression models selected at step 1, 2 and 3 of the procedure, respectively. The estimates of the final model selected at step 4 are shown in Table 2.

It is worth noticing that the estimates $\hat{\beta}_D^{\langle\gamma\rangle}(E)$ are in general very close for the same $D$ and $E$ across Tables 2, 5, 6 and 7. They present some differences because are MLEs computed under different models, but they are similar because they estimate the same quantity. Indeed, as a consequence of the upward compatibility property and of the chosen selection procedure, the parameters $\beta_D^{\langle\gamma\rangle}(E)$ take the same values in the different models considered.

## The two-covariate case

The LML regression model for the distribution of $Y_V|X_{ah}$ is selected using a backward stepwise procedure which, starting from the saturated model, defines a sequence of nested models with zero constraints provided by the set of regression coefficients which are not significatively different from zero. The procedure is based on three ordered steps in which three nested models are fitted, shortly denoted as $M_1$, $M_2$ and $M_3$.

- *step 1*: model $M_1$ with no covariate interactions, i.e. with $\beta_D^{\langle\gamma\rangle}(ah) = 0$ for every $D \subseteq V$, is fitted with a deviance 8.01 on 15 degrees of freedom and $p$-value equal to 0.923. Under $M_1$, some regression coefficients are not significatively different from zero such that the following constraints could be specified for models of $D$-response asssociations of size four,

  - $\beta_D^{\langle\gamma\rangle}(a) = \beta_D^{\langle\gamma\rangle}(h) = 0$, for $D = \{b, c, d, r\}$,

  of size three,

  - $\beta_D^{\langle\gamma\rangle}(a) = 0$, for $D = \{c, d, r\}$,
  - $\beta_D^{\langle\gamma\rangle}(\emptyset) = \beta_D^{\langle\gamma\rangle}(a) = \beta_D^{\langle\gamma\rangle}(h) = 0$, for $D \in \{\{b, c, d\}, \{b, c, r\}, \{b, d, r\}\}$,

  and of size two,

Table 5: LML regression models for $Y_D|X_h$, for every $D \subseteq V$ with $|D| = 1$. The table gives the MLEs of the regression coefficients with their standard error (s.e.) and $p$-value.

| $D$ | $\hat{\beta}_D^{\langle\gamma\rangle}(\emptyset)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ | s.e. | $p$-value |
|---|---|---|---|---|---|---|
| $\{b\}$ | -4.577 | 0.106 | $<.001$ | 2.625 | 0.116 | $<.001$ |
| $\{c\}$ | -4.480 | 0.101 | $<.001$ | 1.056 | 0.143 | $<.001$ |
| $\{d\}$ | -3.256 | 0.054 | $<.001$ | 1.062 | 0.076 | $<.001$ |
| $\{r\}$ | -6.344 | 0.257 | $<.001$ | 3.591 | 0.267 | $<.001$ |

- $\beta_D^{\langle\gamma\rangle}(a) = 0$, for $D \in \{\{b,c\},\{c,r\},\{d,r\}\}$,
- $\beta_D^{\langle\gamma\rangle}(\emptyset) = \beta_D^{\langle\gamma\rangle}(a) = \beta_D^{\langle\gamma\rangle}(h) = 0$, for $D = \{b,d\}$.

- *step 2*: model $M_2$ is fitted including zero constraints for models of $D$-response associations of size four and three shown under $M_1$. The resulting model has a deviance 29.00 on 27 degrees of freedom and $p$-value equal to 0.413. Under $M_2$, regression coefficients which are not significatively different from zero show that the following constraints could be further specified for models of $D$-response associations of size four,

  - $\beta_D^{\langle\gamma\rangle}(\emptyset) = 0$, for $D = \{b,c,d,r\}$,

  and of size two,

  - $\beta_{bc}^{\langle\gamma\rangle}(a) = 0$, for $D \in \{\{b,c\},\{c,r\}\}$,
  - $\beta_{bd}^{\langle\gamma\rangle}(\emptyset) = \beta_{bd}^{\langle\gamma\rangle}(a) = \beta_{bd}^{\langle\gamma\rangle}(h) = 0$, for $D = \{b,d\}$.

- *step 3*: model $M_3$ is fitted including all zero constraints shown under $M_2$. The resulting model has a deviance 33.00 on 33 degrees of freedom and $p$-value equal to 0.467. Regression coefficients under $M_3$ are all significatively different from zero and they are collected in Table 3 of the paper.

Table 6: The selected LML regression models for $Y_D|X_h$, for every $D \subseteq V$ with $|D| = 2$. The table gives the MLEs of the regression coefficients with their standard error (s.e.) and $p$-value, the deviance of the selected models with their degrees of freedom (d.f.) and $p$-value.

| $D$ | $\hat{\beta}_D^{\langle\gamma\rangle}(\emptyset)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ | s.e. | $p$-value | Deviance | d.f. | $p$-value |
|---|---|---|---|---|---|---|---|---|---|
| $\{b\}$ | -4.576 | 0.106 | $<.001$ | 2.623 | 0.115 | $<.001$ | | | |
| $\{c\}$ | -4.479 | 0.101 | $<.001$ | 1.056 | 0.143 | $<.001$ | | | |
| $\{b,c\}$ | 0.472 | 0.181 | 0.009 | $\cdot$ | $\cdot$ | $\cdot$ | 1.103 | 1 | 0.294 |
| $\{b\}$ | -4.576 | 0.106 | $<.001$ | 2.624 | 0.115 | $<.001$ | | | |
| $\{d\}$ | -3.256 | 0.054 | $<.001$ | 1.061 | 0.076 | $<.001$ | | | |
| $\{b,d\}$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | 1.489 | 2 | 0.475 |
| $\{b\}$ | -4.577 | 0.106 | $<.001$ | 2.625 | 0.116 | $<.001$ | | | |
| $\{r\}$ | -6.344 | 0.257 | $<.001$ | 3.592 | 0.266 | $<.001$ | | | |
| $\{b,r\}$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | 0.344 | 2 | 0.842 |
| $\{c\}$ | -4.479 | 0.101 | $<.001$ | 1.056 | 0.143 | $<.001$ | | | |
| $\{d\}$ | -3.256 | 0.054 | $<.001$ | 1.062 | 0.076 | $<.001$ | | | |
| $\{c,d\}$ | 1.773 | 0.182 | $<.001$ | -1.166 | 0.270 | $<.001$ | $\cdot$ | $\cdot$ | |
| $\{c\}$ | -4.480 | 0.101 | $<.001$ | 1.057 | 0.143 | $<.001$ | | | |
| $\{r\}$ | -6.344 | 0.257 | $<.001$ | 3.592 | 0.266 | $<.001$ | | | |
| $\{c,r\}$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | $\cdot$ | 2.946 | 2 | 0.229 |
| $\{d\}$ | -3.255 | 0.054 | $<.001$ | 1.060 | 0.076 | $<.001$ | | | |
| $\{r\}$ | -6.336 | 0.256 | $<.001$ | 3.583 | 0.266 | $<.001$ | | | |
| $\{d,r\}$ | 0.846 | 0.116 | $<.001$ | $\cdot$ | $\cdot$ | $\cdot$ | 1.764 | 1 | 0.184 |

Table 7: The selected LML regression models for $Y_D|X_h$, for every $D \subseteq V$ with $|D| = 3$. The table gives the MLEs of the regression coefficients with their standard error (s.e.) and $p$-value, the deviance of the selected models with their degrees of freedom (d.f.) and $p$-value.

| $D$ | $\hat{\beta}_D^{\langle\gamma\rangle}(\emptyset)$ | s.e. | $p$-value | $\hat{\beta}_D^{\langle\gamma\rangle}(h)$ | s.e. | $p$-value | Deviance | d.f | $p$-value |
|---|---|---|---|---|---|---|---|---|---|
| $\{b\}$ | -4.574 | 0.106 | $<.001$ | 2.621 | 0.115 | $<.001$ | | | |
| $\{c\}$ | -4.477 | 0.101 | $<.001$ | 1.056 | 0.143 | $<.001$ | | | |
| $\{d\}$ | -3.255 | 0.054 | $<.001$ | 1.061 | 0.075 | $<.001$ | | | |
| $\{b,c\}$ | 0.479 | 0.179 | 0.009 | · | · | · | | | |
| $\{b,d\}$ | · | · | · | · | · | · | | | |
| $\{c,d\}$ | 1.775 | 0.182 | $<.001$ | -1.166 | 0.268 | $<.001$ | | | |
| $\{b,c,d\}$ | · | · | · | · | · | · | 3.806 | 5 | 0.578 |
| $\{b\}$ | -4.575 | 0.106 | $<.001$ | 2.620 | 0.115 | $<.001$ | | | |
| $\{c\}$ | -4.478 | 0.101 | $<.001$ | 1.058 | 0.143 | $<.001$ | | | |
| $\{r\}$ | -6.347 | 0.254 | $<.001$ | 3.596 | 0.263 | $<.001$ | | | |
| $\{b,c\}$ | 0.439 | 0.183 | 0.016 | · | · | · | | | |
| $\{b,r\}$ | · | · | · | · | · | · | | | |
| $\{c,r\}$ | · | · | · | · | · | · | | | |
| $\{b,c,r\}$ | 2.910 | 0.345 | $<.001$ | -2.250 | 0.541 | $<.001$ | 4.643 | 5 | 0.461 |
| $\{b\}$ | -4.576 | 0.106 | $<.001$ | 2.624 | 0.115 | $<.001$ | | | |
| $\{d\}$ | -3.255 | 0.054 | $<.001$ | 1.061 | 0.075 | $<.001$ | | | |
| $\{r\}$ | -6.344 | 0.253 | $<.001$ | 3.580 | 0.263 | $<.001$ | | | |
| $\{b,d\}$ | · | · | · | · | · | · | | | |
| $\{b,r\}$ | · | · | · | · | · | · | | | |
| $\{d,r\}$ | 0.853 | 0.115 | $<.001$ | · | · | · | | | |
| $\{b,d,r\}$ | 1.258 | 0.302 | $<.001$ | -0.940 | 0.369 | 0.011 | 3.505 | 5 | 0.623 |
| $\{c\}$ | -4.479 | 0.101 | $<.001$ | 1.058 | 0.143 | $<.001$ | | | |
| $\{d\}$ | -3.255 | 0.054 | $<.001$ | 1.058 | 0.076 | $<.001$ | | | |
| $\{r\}$ | -6.336 | 0.254 | $<.001$ | 3.584 | 0.264 | $<.001$ | | | |
| $\{c,d\}$ | 1.775 | 0.181 | $<.001$ | -1.233 | 0.272 | $<.001$ | | | |
| $\{c,r\}$ | · | · | · | · | · | · | | | |
| $\{d,r\}$ | 0.831 | 0.117 | $<.001$ | · | · | · | | | |
| $\{c,d,r\}$ | · | · | · | · | · | · | 5.039 | 5 | 0.411 |

# References


Agresti, A. (2013). *Categorical data analysis, 3rd edition*. John Wiley & Sons.

Bartolucci, F., R. Colombi, and A. Forcina (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica 17*, 691–711.

Bergsma, W., M. Croon, and J. Hagenaars (2009). *Marginal models for dependent, clustered, and longitudinal categorical data*. London, UK: Springer.

Bergsma, W. P. and T. Rudas (2002). Marginal log-linear models for categorical data. *Annals of Statistics 30*(1), 140–159.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics 41*, 691–711.

Berrington de González, A. and D. R. Cox (2007). Interpretation of interaction: A review. *The Annals of Applied Statistics 1*(2), 371–385.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B 41*(1), 1–31.

Deeks, S. G. and A. N. Phillips (2009). HIV infection, antiretroviral treatment, ageing, and non-AIDS related morbidity. *BMJ: British Medical Journal 338*(7689), pp. 288–292.

Drton, M. (2009). Discrete chain graph models. *Bernoulli 15*(3), 736–753.

Drton, M. and T. Richardson (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B 70*(2), 287–309.

Ekholm, A., J. W. McDonald, and P. W. F. Smith (2000). Association models for a multivariate binary response. *Biometrics 56*(3), 712–718.

Evans, R. and A. Forcina (2013). Two algorithms for fitting constrained marginal models. *Computational Statistics and Data Analysis 66*, 1–7.

Glonek, G. J. N. and P. McCullagh (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B 57*(3), 533–546.

Guaraldi, G., G. Orlando, S. Zona, M. Menozzi, F. Carli, E. Garlassi, A. Berti, E. Rossi, A. Roverato, and F. Palella (2011). Premature age-related comorbidities among hiv-infected persons compared with the general population. *Clinical Infectious Diseases 53*(11), 1120–1126.

Lang, J. and A. Agresti (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association 89*(426), 625–632.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *The Annals of Statistics 24*(2), 726–752.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Marchetti, G. M. and M. Lupparelli (2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli 17*(3), 827–844.

Marengoni, A., S. Angleman, R. Melis, F. Mangialasche, A. Karp, A. Garmen, B. Meinow, and L. Fratiglioni (2011). Aging with multimorbidity: A systematic review of the literature. *Ageing research reviews 10*(4), 430–439.

May, M. T. and S. M. Ingle (2011). Life expectancy of HIV-positive adults: a review. *Sexual Health 8*(4), 526–533.

McCullagh, P. and A. Nelder, J. (1989). *Generalized linear models, 2nd edition*. London, UK: Chapman and Hall.

Mocroft, A., B. Ledergerber, C. Katlama, O. Kirk, P. Reiss, A. d'Arminio Monforte, B. Knysz, M. Dietrich, A. Phillips, and J. Lundgren (2003). Decline in the AIDS and death rates in the EuroSIDA study: an observational study. *The Lancet 362*(9377), 22–29.

Molenberghs, G. and E. Lesaffre (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine 18*, 2237–2255.

Phillips, A. N., J. Neaton, and J. D. Lundgren (2008). The role of HIV in serious diseases other than AIDS. *AIDS 22*(18), 2409–2418.

Roverato, A. (2015). Log-mean linear parameterization for discrete graphical models of marginal independence and the analysis of dichotomizations. *Scandinavian Journal of Statistics 42*(2), 627–648.

Roverato, A., M. Lupparelli, and L. La Rocca (2013). Log-mean linear models for binary data. *Biometrika 100*, 485–494.

Shiels, M. S., R. M. Pfeiffer, and E. A. Engels (2010). Age at cancer diagnosis among persons with aids in the united states. *Annals of Internal Medicine 153*(7), 452–460.

Tutz, G. (2011). *Regression for categorical data*. Cambridge University Press.

Wermuth, N. and K. Sadeghi (2012). Sequences of regressions and their independences. *TEST 21*(2), 215–252.