Journal of Statistical Planning and Inference 143 (2013) 63–81

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

The asymptotic covariance matrix of the odds ratio parameter estimator in semiparametric log-bilinear odds ratio models

Angelika Franke, Gerhard Osius *

Faculty 3, Mathematics/Computer Science, University of Bremen, Germany

ARTICLE INFO

Article history:
Received 5 May 2011
Received in revised form
12 April 2012
Accepted 5 June 2012
Available online 15 June 2012

Keywords:
Odds ratio
Asymptotic
Covariance matrix
Conditional sampling
Semiparametric
Log-linear models
Log-bilinear association
Logistic regression
Linear regression

ABSTRACT

The association between two random variables is often of primary interest in statistical research. In this paper semiparametric models for the association between random vectors X and Y are considered which leave the marginal distributions arbitrary. Given that the odds ratio function comprises the whole information about the association, the focus is on bilinear log-odds ratio models and in particular on the odds ratio parameter vector θ. The covariance structure of the maximum likelihood estimator $\hat{\theta}$ of θ is of major importance for asymptotic inference. To this end different representations of the estimated covariance matrix are derived for conditional and unconditional sampling schemes and different asymptotic approaches depending on whether X and/or Y has finite or arbitrary support. The main result is the invariance of the estimated asymptotic covariance matrix of $\hat{\theta}$ with respect to all above approaches. As applications we compute the asymptotic power for tests of linear hypotheses about θ—with emphasis to logistic and linear regression models—which allows to determine the necessary sample size to achieve a wanted power.

1. Introduction and outline

The question how a random output vector Y of a system (e.g. the health status of a human) is associated to a random input vector X (e.g. consumption of tobacco and alcohol, environmental pollution and other risk factors) is of major importance in statistical science. If the association between X and Y is of primary interest, then a semi-parametric association is appropriate which leaves the marginal distributions of X and Y arbitrary. However, the association is completely determined by the odds-ratio function $OR(x,y)$ for the joint density $p(x,y)$ with respect to fixed reference values x_0 and y_0 (cf. Osius, 2004, 2009):

$$OR(x,y) = \frac{p(x,y) \cdot p(x_0,y_0)}{p(x,y_0) \cdot p(x_0,y)}. \tag{1.1}$$

A semi-parametric odds-ratio model specifies this function up to an unknown parameter vector θ, but leaves marginal distributions arbitrary. An important class are log-bilinear odds-ratio models given by

$$\log OR(x,y) = \tilde{x}^T \theta \tilde{y}, \tag{1.2}$$

where $\tilde{x}$ and $\tilde{y}$ are known vector-valued functions of x and y which may coincide with x and y, respectively. The association structure of some widely used regression models is log-bilinear, e.g. generalized linear models with canonical link (for

* Corresponding author.
E-mail address: osius@math.uni-bremen.de (G. Osius).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.06.005

univariate Y), multivariate linear logistic regression (for Y with finite support) and multivariate linear regression. An advantage of odds-ratio models over these regression models is that inference about the association parameter θ may also be obtained from samples drawn conditionally on Y (instead of X). Generalizing an important result by Prentice and Pyke (1979), it has been shown in Osius (2009) that the estimator $\hat{\theta}$ and its estimated asymptotic covariance matrix $Cov_\infty(\hat{\theta})$ for samples conditional on Y are exactly the same as if the sample had been drawn conditionally on X. The purpose of this paper is to derive different representations of this covariance matrix on which statistical analysis (e.g. tests and confidence regions) is based. These results are applied to compute the asymptotic power for tests of linear hypothesis about θ which allows to determine the sample sizes necessary to achieve a wanted power.

A given random sample $(X_i,Y_i), i=1,\ldots,n$ containing $J+1$ different X-values $X_{(0)},\ldots,X_{(J)}$ and $K+1$ different Y-values $Y_{(0)},\ldots,Y_{(K)}$ can be summarized by the counts

$$R_{jk} = \#\{i \mid X_i = X_{(j)},\ Y_i = Y_{(k)}\} \tag{1.3}$$

for the observed combinations (j,k). Although the distribution of the table (R_{jk}) depends on the sampling scheme (e.g. conditional on X or Y), we will show that the estimated asymptotic covariance matrix of $\hat{\theta}$ is invariant against common sampling schemes and asymptotic approaches. However, we do not establish original asymptotic results here but—using mainly matrix algebra—derive different representations for asymptotic covariance matrices and in particular for $Cov_\infty(\hat{\theta})$.

The paper is organized as follows. Section 2 gives a brief introduction to odds ratio models with emphasis on multivariate linear logistic regression (where Y has finite support) and log-linear models for contingency tables (where the support of X is finite too). The next Section 3 deals with estimation of θ under different sampling schemes (unconditional and conditional on X and Y, respectively). Our main results are contained in Section 4. Based on the work of Haberman (1974) we first show that for contingency tables (i.e. both X and Y have finite support) the asymptotic distribution of $\hat{\theta}$ is invariant under the common sampling schemes and provide different representations of $Cov_\infty(\hat{\theta})$. Looking more generally at the multivariate linear logistic regression model (with arbitrary support of X) and sampling conditional on X we observe that the estimated asymptotic covariance matrix $Cov_\infty(\hat{\theta})$ is the same as for contingency tables (where X has finite support). The general case allowing arbitrary supports for X and Y is dealt with in Section 5. For sampling conditional on Y and a fixed set of conditioning values we conclude that the matrix of $Cov_\infty(\hat{\theta})$ is the same as before where both X and Y had finite support. As a first application we show in Section 6 how our results can be used to compute the asymptotic power for testing a linear hypothesis $Q\theta = 0$ and how to determine the necessary sample size to achieve a given power for a value θ' of interest under the alternative. Finally, we demonstrate for univariate Y how the linear resp. log-linear model emerges from an odds-ratio-model by imposing additional assumptions on the conditional distribution of Y (given X) and conclude with a short discussion of our results. In Appendices A and B we summarize some definitions and results from linear algebra, which are used freely throughout the paper without explicit reference. In Appendix C the proofs of the theorems are given.

2. Odds ratio models

Consider a pair (X,Y) of random vectors defined on some probability space taking values in $\Omega = \Omega_X \times \Omega_Y \subset \mathbb{R}^{M_X} \times \mathbb{R}^{M_Y}$ with joint distribution P and marginal distributions P^X and P^Y. To avoid trivialities we assume that Ω_X and Ω_Y both have more than one element. Let ν_X and ν_Y be two fixed σ-finite measures on $\mathbb{R}^{M_X}$ and $\mathbb{R}^{M_Y}$ such that P has a positive density p on Ω with respect to the product measure $\nu = \nu_X \times \nu_Y$—typically a product of Lebesgue or counting measures. The log-density can be parametrized as

$$\log p(x,y) = \alpha + \rho(x) + \gamma(y) + \psi_\theta(x,y), \quad x \in \Omega_X,\ y \in \Omega_Y \tag{2.1}$$

with integrable functions ρ, γ, ψ, an unknown parameter $\theta \in \Theta$, and an integration constant α determined by $\int p\, d\nu = 1$. To guarantee identifiability we assume the constraints

$$\rho(x_0) = \gamma(y_0) = 0, \tag{2.2}$$

where $x_0 \in \Omega_X$ and $y_0 \in \Omega_Y$ are the reference values of the odds ratio function. The conditional distribution of Y given X has a positive density $p(y \mid X = x)$ given by

$$\log p(y \mid X = x) = \gamma(y) + \psi_\theta(x,y) - \delta_\theta(x) \tag{2.3}$$

with an integration constant $\delta_\theta(x)$ and similarly

$$\log p(x \mid Y = y) = \rho(x) + \psi_\theta(x,y) - \varepsilon_\theta(y). \tag{2.4}$$

An important class of parametric association models are log-bilinear association models with respect to the transformed variables $\tilde{x} = h_X(x)$ and $\tilde{y} = h_Y(y)$ given by measurable maps $h_X : \mathbb{R}^{M_X} \to \mathbb{R}^{L_X}$ and $h_Y : \mathbb{R}^{M_Y} \to \mathbb{R}^{L_Y}$ which will always be chosen here such that $\tilde{x}_0 = h_X(x_0) = 0$ and $\tilde{y}_0 = h_Y(y_0) = 0$. The functions h_X and h_Y are typically injective (one-to-one) but to avoid trivialities we merely assume that they are not constant. The parameter θ is a $L_X \times L_Y$-matrix and the log-odds ratio function is bilinear in the transformed variables $\tilde{x}$ and $\tilde{y}$

$$\psi_\theta(x,y) = \tilde{x}^T \theta \tilde{y} \quad \text{for all } x,y. \tag{2.5}$$