ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Joint regression analysis of mixed-type outcome data via efficient scores

Scott Marchese, Guoqing Diao *

George Mason University, Fairfax VA, USA

ARTICLE INFO

Article history:
Received 3 August 2017
Received in revised form 28 February 2018
Accepted 28 February 2018
Available online xxxx

Keywords:
Bonferroni correction
Efficient score
Generalized estimating equations
Mixed-type data
Multiplier bootstrap

ABSTRACT

Joint analysis of multivariate outcomes composed of mixed data types (continuous, count, binary, survival, etc.) induces special complexity in model specification and analysis. When the scientific question of interest involves a joint effect of covariate(s) of interest on the set of outcome variables, specifying a full probability model may be infeasible, undesirably complex, or computationally intractable. A flexible method to estimate and conduct inference on such joint effects is presented which accounts for correlation among the outcomes without needing to explicitly specify their joint distribution. Simulation studies and an analysis of health care data illustrate the approach and its operating characteristics vis-à-vis other methods.

1. Introduction

Multivariate outcomes arise naturally in a variety of data analysis contexts: in health studies with multiple indicators of disease or wellness; in pharmacology, when various properties of a drug need to be controlled for safety, efficacy, etc., or various biomarkers are profiled to form a multifarious description of cellular activity; in survey sampling where multiple aspects of a topic or product are of inherent interest, and so on. Many subdomains of statistics address specific types of multivariate or correlated outcomes, such as longitudinal data analysis which provides various tools for data that may be identically distributed on the margins (up to covariate-dependent scale/location shifts) but not independent. Many regression approaches assume, in some form, that the data were generated from a multivariate normal distribution (e.g. linear mixed models and latent variable methods) or employ an approach such as estimating equations which model the mean function and dependence parameters separately without attempting to specify a full probability model for the data (generalized linear mixed models or GLMM).

Copulas provide another attractive approach to modeling the joint distribution of a variety of outcomes (Song, 2007), though they are less straightforward to implement when some margins are not continuous. With discrete – or mixed continuous and discrete – margins, subtle identifiability issues complicate model specification and introduce considerations which may be overly technical for a data analyst (for a discussion see, e.g., Genest and Nešlehová, 2007). Although a wide variety of copula models exist and have some degree of user-friendly implementation of the copula itself, accessible implementation of such models for regression analysis remains relatively sparse.

For some common cases of mixed-type outcomes a number of models have been developed: with one continuous and one binary outcome, for instance, various factorization or latent variable methods exist (Teixeira-Pinto and Harezlak, 2013), although the 'direction' of factorization (specifying $f(Y_1, Y_2) = f(Y_1) \cdot f(Y_2|Y_1)$ or $f(Y_1, Y_2) = f(Y_2) \cdot f(Y_1|Y_2)$) generally yields

* Correspondence to: Department of Statistics, George Mason University, 4400 University Drive, MS 4A7, Fairfax, VA 22030, USA.
E-mail addresses: smarches@masonlive.gmu.edu (S. Marchese), gdiao@gmu.edu (G. Diao).

https://doi.org/10.1016/j.csda.2018.02.008
0167-9473/

distinct model interpretations and results. Such models as appear in domain-specific applications, for example the genome-wide association study literature (Kwak et al., 2013), often make quite restrictive assumptions about the types of data involved. In this context, one inferential goal is a test of the effect of genetic information on disease-related phenotypes, but the need to specify a coherent multivariate distribution for these outcomes may induce considerable rigidity of assumptions.

Despite such shortcomings, models for multivariate outcomes have the attractive feature of allowing hypothesis tests of the joint effect of covariate(s) of interest on the set of outcomes. Such a test may be particularly useful in low-power settings, whether due to small n or inherently noisy data, or in cases where sets of outcomes form natural groupings (ex. groups of related biomarkers). While a joint model for outcomes may increase power for marginal tests of significance, the magnitude of the gain is often small or even nonexistent (Teixeira-Pinto and Normand, 2009; Teixeira-Pinto and Harezlak, 2013). In addition to fitting regression parameters, multivariate models often provide some information about the residual covariance of the outcomes. Although potentially useful, such information may be of limited application or even misleading due to (i) model mis-specification, which seems much more likely in the multivariate context, or (ii) use of 'robust' (sandwich) covariance estimators which are asymptotically consistent for the regression parameters (and linear combinations thereof) but do not presume a correct model for the correlation parameters. In any case, such correlation parameters are usually of secondary interest — indeed, they are often relegated to 'nuisance' status.

This paper presents a relatively straightforward strategy for testing the joint effect of a covariate of interest on a multivariate vector of outcomes, or for testing more general linear combinations of regression parameters across different outcomes. The procedure outlined employs a so-called *multiplier bootstrap* method which accounts for the correlation between outcomes without needing to explicitly model it. So long as a regression model may be fit separately for each margin (and given some mild regularity conditions), the method provides valid hypothesis tests for a wide variety of joint hypotheses. This test is 'parameterized' by a norm, and different choices of norms affect the sensitivity of the test toward different types of alternative hypotheses. In particular, the relative sparsity or density of non-zero coefficients among the set of tested parameters corresponds in an intuitive way to the norm chosen in the bootstrap procedure.

Since many extant models provide ways of effectively analyzing (i) multivariate, continuous data (copulas) or (ii) data which arise from the same family of marginal distributions but which are not independent (such as GLMM), the present model is geared toward instances of dependent *and* mixed-type data. Naturally, the method applies to the two aforementioned situations but might answer a less comprehensive set of statistical questions than other, more tailored models in those contexts.

2. Methods

2.1. Background

The proposed procedure takes its cues from some methods developed in the high-dimensional data context. Consider the realm of genomics, where genetic information is measured from a (typically relatively small) sample of patients, animals, or cells. Sequencing technology allows collection of various measurements including quantitative trait loci, single nucleotide polymorphisms, RNA expression levels, knockout–induced pathway activity modulation, etc. In all cases, a similar problem arises: the same model is fit at many locations (on many small data sets), and the resulting test statistics usually exhibit non-ignorable dependence, making 'genome-wide' or multiple-testing corrected cutoff regions difficult to calibrate. Earlier literature (ex. Rebai et al., 1994) focused on employing approximations to functions of such correlated test statistics (for example, their maximum); however, this approach was often model-specific and depended critically on simplifying distributional assumptions.

To overcome such limitations, permutation and bootstrap approaches gained favor: for analysis of quantitative trait loci data, Zou et al. (2004) implement a bootstrap test for genome-wide statistical significance testing, employing standard normal perturbations of score information to estimate quantiles of a random variable which defines the rejection region. Diao and Vidyashankar (2013) modify this procedure by replacing the normal perturbations with Rademacher ones, and Diao et al. (2014) apply this so-called multiplier bootstrap procedure more generally in the context of $p \gg n$ linear regression problems. Therein, p covariates $\mathbf{x} \equiv [x_1, \ldots, x_p]$ model a univariate outcome Y: $E(Y|\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$. In this setting, the covariance matrix of $\boldsymbol{\beta}$ is not estimable since, due to sample size constraints, $\mathbf{X}^\top \mathbf{X}$ is not of rank p.

Various methods exist to attain a modified, usable estimate of the covariance matrix, in particular via shrinkage of eigenvalues (Ledoit and Wolf, 2012). In Kuelbs and Vidyashankar (2010), asymptotic theory for this application is developed and supplemented with some hypothesis testing examples wherein the supremum norm of a Gaussian process defines the rejection region for the global hypothesis $H_0 : \boldsymbol{\beta} = 0$ when the dimension of $\boldsymbol{\beta}$ tends to ∞ along with n. They use a bootstrap test which generates random vectors based on the shrunken covariance estimate – although theoretically promising, this procedure's computational complexity scales quite poorly with p. The multiplier bootstrap procedure in Diao et al. (2014) modifies this test by using a Rademacher process to perturb the efficient scores, thereby generating a null distribution of the test statistic which reflects correlation among the parameters in $\boldsymbol{\beta}$ without needing to explicitly model said correlation. The authors demonstrate that this multiplier test may easily handle large p without imposing an excessive computational burden.

The problem addressed here arises in a different data structure but encounters a similar problem: in the $p \gg n$ case, an estimate of the covariance matrix is straightforward in principle, but due to practical constraints (sample size),

Please cite this article in press as: Marchese S., Diao G., Joint regression analysis of mixed-type outcome data via efficient scores. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.02.008.