SDS
Statistica e
Data Science

Palermo, 11-12 April 2024

Proceedings of the Statistics and Data Science 2024 Conference

New perspectives on Statistics and Data Science

Edited by

Antonella Plaia – Leonardo Egidi
Antonino Abbruzzo

Proceedings of the SDS 2024 Conference
Palermo, 11-12 April 2024
Edited by: Antonella Plaia – Leonardo Egidi - Antonino Abbruzzo
–
Palermo: Università degli Studi di Palermo.

ISBN Ebook (Pdf)
978-88-5509-645-4

A cluster-weighted model for COVID-19 hospital admissions

Daniele Spinelli, Paolo Berta, Salvatore Ingrassia and Giorgio Vittadini

Abstract We propose a cluster-weighted model to analyze the mortality and the latent heterogeneity of COVID-19 patients. We focus on administrative data collected during in the earliest phases of the COVID-19 pandemic. Results highlight that a model-based clustering approach is helpful to detect unobserved clusters of COVID-19 patients.

Key words: Cluster-Weighted Models, COVID-19, clustering, administrative data

1 Introduction

Several studies have focused on the characteristics underlying COVID-19 mortality. Contributing causes of death by COVID-19 have been investigated using different statistical approaches to highlight the effect on mortality of age, gender, comorbidities, time of onset of illness, and other characteristics. What seems to dominate from a statistical point of view, is the large heterogeneity of the populations affected by COVID-19 and the difficulty in identifying subpopulations when classical statistical

Daniele Spinelli
Department of Statistics and Quantitative Methods, University of Milano-Bicocca e-mail: daniele.spinelli@unimib.it

Paolo Berta
Department of Statistics and Quantitative Methods, University of Milano-Bicocca e-mail: paolo.berta@unimib.it

Salvatore Ingrassia
Department of Economics and Business , University of Catania e-mail: salvatore.ingrassia@unict.it

Giorgio Vittadini
Department of Statistics and Quantitative Methods, University of Milano-Bicocca e-mail: giorgio.vittadini@unimib.it

approaches, such as generalized linear models, are adopted. The most common empirical strategies assume that the population affected by COVID-19 is homogeneous. However, some studies have considered latent heterogeneity in relation to COVID-19 [3, 6, 8]. These studies have also used clustering methods to reveal unobserved subpopulations in COVID-19 patients. The first study [3] used unsupervised clustering techniques on clinical data to identify three subgroups of pediatric patients and identified the characteristics of a multisystem inflammatory syndrome in children phenotype. Moreover, [6] used unsupervised cluster analysis on Spanish hospital records to identify subgroups of patients exhibiting post-COVID symptoms. The obtained clusters are to be used to assign therapeutic interventions. Further, [8] used a two-step method combining clustering and generalized linear models. In the first step, they used a k-means algorithm to reveal clusters of COVID-19 patients on a longitudinal dataset of biomarkers measurements. In the second stage, they used the obtained clusters as covariates to predict mortality using logistic regression. Their purpose is similar to ours, as they relate the clustering to mortality. However, we employ a different modeling technique.

Specifically, we propose a clustering framework to study COVID-19 mortality adopting a cluster-weighted model (CWM; [4, 5]). Compared to the above mentioned unsupervised methods, this supervised method has the advantage to model the latent heterogeneity, the outcomes of interest and the covariates affecting the outcome simultaneously. This also means, in contrast to the estimating procedure used in [8], that the effects of the covariates on mortality is allowed to vary within each cluster instead of being assumed independent of the latent class structure. Indeed, the CWM does not rely on the assignment independence assumption which states that the assignment of the data points to the latent clusters is independent from the covariates distribution [7] .

2 Data

We consider administrative data regarding 2,617 Covid-19 hospitalizations occurred in the period from January to June 2020 at the Spedali–Civili hospital in Brescia, Italy [1]. The data include demographic characteristics, clinical information (diagnoses according to ICD-9-CM classification) and admission characteristics such as admission date, discharge date, in-hospital mortality, unit of admission and procedures performed during the hospitalization.

3 Methods

We assume a sample $(x_1, y_1), \ldots, (x_n, y_n)$ concerning a response variable Y and a set of covariates X. The sample is hypothesized to come from a heterogeneous population formed by K latent classes. The CWM models the density of (Y, X) as outlined

by Eq. 1

$$p(x,y,\theta) = \sum_{j=1}^{K} \pi_j d(y|x;\beta_j) q(x;\eta_j). \quad (1)$$

In Eq. 1, π_j is the mixing proportion of latent class j, $d(y|x;\zeta_j)$ is the class j-specific conditional density of the response variable and $q(x;\eta_j)$ is the marginal density of X in class j. Parameter vectors β_j and η_j are included in the generic vector of parameters θ to be estimated for all $j = 1,...,k$. Specifically, β_j is a vector of regression coefficients, and η_j depends on distributional assumptions regarding the covariates. Assuming $d(y|x;\beta_j) = 1$ in Eq. 1 leads to a mixture of distributions, while $q(x;\eta_j) = 1$ leads to a finite mixture of regressions (FMR).

Our response variable Y is a binary mortality indicator. We assume that Y follows a Bernoulli distribution with probability function $f(x;\beta_j)$, where x is a set of categorical and numeric risk-adjustment covariates, $f(.)$ is the logistic function and β_j is a set of latent class-specific regression coefficients to be estimated. Therefore, $d(y|x;\beta_j)$, the conditional density of Y for latent class j (Eq. 1) is as follows:

$$d(y|x;\beta_j) = \Big[f(x;\beta_j)\Big]^{y} \Big[1 - f(x;\beta_j)\Big]^{1-y}. \quad (2)$$

In Eq. 1, the conditional density $q(x;\eta_j)$ depends on distributional assumptions on the covariates. The vector of covariates X includes age (continuous), sex (dichotomous), number of comorbidities (discrete) and period of admission in weeks from the start of 2020 (continuous). Patient characteristics are included to control for clinical pre-existing conditions, and they can be considered as a risk factor for in-hospital mortality. The period of admission is included as a proxy for the stress experienced by the healthcare system.

Period of admission and age are modeled according to a multivariate Gaussian distribution with mean vector μ_j and variance-covariance matrix Σ_j, considering Gaussian parsimonious models [2]. As for the number of comorbidities, we assume a Poisson distribution with mixture component-specific mean λ_j, while sex is assumed to be Bernoulli distributed with probability ψ_j. Hence, in keeping with the notation in Eq. 1, the expression for the marginal density of the covariates is the following:

$$\begin{aligned} q(x;\eta_j) &= q(x;\mu_j,\Sigma_j,\lambda_j,\psi_j) = \\ &= \phi(Age, Period;\mu_j,\Sigma_j) \times \frac{\lambda_j^{Comorb} e^{-\lambda_j}}{Comorb!} \times (\psi_j)^{Sex} (1-\psi_j)^{1-Sex} \end{aligned} \quad (3)$$

where, $\phi(.)$ is the bivariate Gaussian probability density function. The eigenvalue decomposition of Σ_j leads to 14 possible parametrizations. The model obtained by combining Eqs. 1, 2 and 3 is estimated in Stata by the EM algorithm implemented in the `cwmglm` package [9]. We consider competing CWMs and FMRs with varying number of latent clusters and different parsimonious model specification on the multivariate Gaussian covariates and deselect them using the Akaike information

criterion (AIC) and Bayesian information criterion (BIC). We cluster observations according to their maximum a posteriori probability.

4 Results

Our preferred model , a CWM with $k = 3$, is used to identify different profiles in terms of covariates (age, sex, period of admission and comorbidities) and mortality risks. The first group ($\pi_1 = 0.07$) is composed mainly by the oldest people (average age 74 years), who have been hospitalized later, on average. Patients in group 2 ($\pi_2 = 0.79$) are on average aged 70 years old and have been admitted earlier. Group 3 ($\pi_3 = 0.14$) is characterized by the lowest mean age (46) and by the lowest mortality. The impact of covariates on mortality differs from what obtained in [8], where the chosen statistical method could only estimate a single odds ratio for age and sex after controlling for clustering assignment. Our modeling framework is more flexible as the risk related to the covariates is allowed to be cluster-dependent. Consistent with [8], our estimation reveals a positive effect of age on mortality in all of the clusters, but the magnitude is different. The odds ratio are 1.04, 1.10 and 1.24; their estimate is 1.08 after controlling for cluster membership.

References

[1] Berta, P., Ingrassia, S., Vittadini, G., Spinelli, D.: Latent heterogeneity in covid-19 hospitalisations: a cluster-weighted approach to analyse mortality. Aust NZ J Stat (2024)

[2] Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**(5), 781–793 (1995)

[3] Geva, A., Patel, M.M., Newhams, M.M., Young, C.C., Son, M.B.F., Kong, M., Maddux, A.B., Hall, M.W., Riggs, B.J., Singh, A.R., et al.: Data-driven clustering identifies features distinguishing multisystem inflammatory syndrome from acute covid-19 in children and adolescents. EClinicalMedicine **40** (2021)

[4] Ingrassia, S., Minotti, S., Vittadini, G.: Local statistical modeling via the cluster-weighted approach with elliptical distributions. J. Classif. **29**(3), 363–401 (2012)

[5] Ingrassia, S., Punzo, A., Vittadini, G., Minotti, S.: The generalized linear mixed cluster-weighted model. J. Classif. **32**(1), 85–113 (2015)

[6] Fernández-de Las-Peñas, C., Martín-Guerrero, J.D., Florencio, L.L., Navarro-Pardo, E., Rodríguez-Jiménez, J., Torres-Macho, J., Pellicer-Valero, O.J.: Clustering analysis reveals different profiles associating long-term post-covid symptoms, covid-19 symptoms at hospital admission and previous medical comorbidities in previously hospitalized covid-19 survivors. Infection **51**(1), 61–69 (2023)