IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment

BOOK OF SHORT PAPERS

Editors: Rosaria Lombardo, Ida Camminatiello and Violetta Simonacci

Book of Short papers
10th International Conference **IES 2022**
Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment

Department of Economics, University of Campania "L. Vanvitelli",
January 27th - 28th 2022

Scientific Committee of the group of the Italian Statistical Society on Statistics for the Evaluation and Quality of Services – SVQS

Pietro Amenta -University of Sannio
Matilde Bini -European University of Roma
Luigi D'Ambra -University of Naples "Federico II"
Maurizio Carpita -University of Brescia
Paolo Mariani -University of Milan "Bicocca"
Marica Manisera -University of Brescia
Monica Palma -University of Salento
Pasquale Sarnacchiaro -University of Rome Unitelma Sapienza

Program Committee of the conference IES 2022

Chair: Rosaria Lombardo -University of Campania "L. Vanvitelli"
Fabio Bacchini -ISTAT
Laura Baraldi -University of Campania "L. Vanvitelli"
Eric Beh -University of Newcastle, Australia
Wicher Bergsma -The London School of Economic and Political Science, UK
Enrico Bonetti -University of Campania "L. Vanvitelli"
Eugenio Brentari -University of Brescia
Clelia Buccico -University of Campania "L. Vanvitelli"
Rosalia Castellano -University of Naples "Parthenope"
Ida Camminatiello -University of Campania "L. Vanvitelli"
Carlo Cavicchia -University of Rotterdam, The Netherlands
Enrico Ciavolino -University of Salento
Corrado Crocetta -University of Foggia
Claudio Conversano -University of Cagliari
Antonello D'Ambra -University of Campania "L. Vanvitelli"
Antonio D'Ambrosio -University of Naples "Federico II"
Alfonso Iodice D'Enza -University of Naples "Federico II"
Tonio Di Battista -University of Chieti "G. D'Annunzio"
Michele Gallo -University of Naples "L'Orientale"
Francesco Gangi -University of Campania "L. Vanvitelli"
Michele La Rocca -University of Salerno
Amedeo Lepore -University of Campania "L. Vanvitelli"
Riccardo Macchioni -University of Campania "L. Vanvitelli"
Filomena Maggino -University of Rome "La Sapienza"
Angelos Markos -Demokritus University of Thrace, Greece
Lucio Masserini -University of Pisa
Stefania Mignani -University of Bologna
Nicola Moscariello -University of Campania "L. Vanvitelli"
Francesco Palumbo -University of Naples "Federico II"
Alessandra Petrucci -University of Florence
Alessio Pollice -University of Bari

Donato Posa -University of Salento
Luca Secondi -University of Tuscia
Amalia Vanacore -University of Naples "Federico II"
Michel van de Velden -University of Rotterdam, The Netherlands
Rosanna Verde -University of Campania "L. Vanvitelli"
Donatella Vicari -University of Rome "La Sapienza"
Grazia Vicario -Polytechnic University of Turin
Maurizio Vichi -University of Rome "La Sapienza"

Organizing Committee

Ida Camminatiello -University of Campania "L. Vanvitelli"
Antonello D'Ambra -University of Campania "L. Vanvitelli"
Rosaria Lombardo -University of Campania "L. Vanvitelli"
Elvira Romano -University of Campania "L. Vanvitelli"
Luca Rossi -University Cusano
Violetta Simonacci -University of Naples "L'Orientale"

Editors

Rosaria Lombardo - University of Campania “L. Vanvitelli”, Italy
Ida Camminatiello - University of Campania “L. Vanvitelli”, Italy
Violetta Simonacci - University of Naples “L’Orientale”, Italy

PKE - Professional Knowledge Empowerment s.r.l.
Sede legale: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Sede operativa: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Ufficio Di Rappresentanza: Via Giacomo Peroni, 400 - 00131 Roma (RM)
CF / P.I. 03167830920 — www.pke.it; e-mail info@pke.it — Privacy

February 2022 PKE s.r.l.
ISBN 978-88-94593-35-8 on print
ISBN 978-88-94593-36-5 online

PKE Publisher

Contents

A Bayesian non parametric approach for bias correction for underreported data.

Un approccio Bayesiano non parametrico per la correzione dei dati sotto riportati.

Serena Arima, Giuseppe Pasculli and Silvia Polettini

Abstract Data quality is emerging as an essential characteristics of all data driven processes. The implications that data quality issues in computing health or vital statistics may have on government intervention policies and distribution of financial resources highlight the relevance of the problem. In this paper, we deal with the issue of underreporting, paying particular attention to its effects on the estimation of the prevalence of a phenomenon. We propose a non parametric compound Poisson model that allows for the estimation of the reporting probabilities. The proposed model will be applied to a data set concerning early neonatal mortality in Minas Gerais, Brazil. Comparisons of the estimates obtained under several alternative models reveal that the proposed approach is accurate and particularly suitable when there is no prior information about the reporting probability.

Abstract Negli ultimi anni e durante la pandemia, si é confermato il fatto che i dati di qualità sono l'elemento fondamentale per tutti i processi di decisione basati sui dati. Infatti, dati collezionati in modo viziato portano a decisioni distorte con conseguenze su azioni politiche e distribuzioni di fondi. Ció è particolarmente importante se si tratta di sanitá e salute pubblica. In questo lavoro, consideriamo il problema della distorsione dei dati con particolare interesse agli effetti che questo ha sulla stima della prevalenza di un fenomeno. Proponiamo un modello non parametrico di Poisson per la stima delle probabilitá di tale distorsione. Il modello è applicato a dati relativi alla mortalitńeonatale in Minas Gerais, Brasile. Le stime verranno confrontante con modelli alternativi e l'applicazione evidenzia che il metodo

Serena Arima
Departiment of history, society and social sciences, University of Salento, Lecce, Italy, e-mail: serena.arima@unisalento.it;
Giuseppe Pasculli
Department of Computer, Control, and Management Engineering Antonio Ruberti (DIAG), University of Rome "La Sapienza", Rome, Italy, e-mail: giuseppe.pasculli@uniroma1.it ;
Silvia Polettini
Department of Political Sciences, University of Rome "La Sapienza", Rome, Italy, e-mail: silvia.polettini@uniroma1.it

proposto é particolarmente promettente laddove non ci sono informazioni a-priori circa la qualità del dato raccolto.

Key words: Compund Poisson model, hierarchical models, MCMC, Underreporting probabilities,

1 Introduction

In order for scientists to derive reliable statistical analyses, useful for taking appropriate data driven decisions, good quality data are mandatory. As experienced in the last year, inaccurate data collection leads to inappropriate conclusions even when accurate and complex statistical methodologies had been performed. The problem is particularly severe when health or vital statistics are concerned, with important consequences on government intervention policies and distribution of financial resources. In underdeveloped and developing countries, the poor quality of the available data often impairs estimation of economic, health, and social indicators. An example is provided by the case study of this paper, originally presented in [4], where early neonatal mortality risk in Minas Gerais, Brazil, during 1999-2001 is studied. The state of Minas Gerais presents heterogeneous characteristics and a relevant socio-economical inequality. In this context, it is expected that official figures do not reflect the real situation with reports most likely overlooking deaths occurring soon after birth. If such phenomena are not taken into proper account, critical statistics may be underestimated, jeopardizing the concept of effective government intervention policies and the allocation of financial resources.
The problem of underreporting is well known in the literature. The bias caused by a flawed data reporting method is widely discussed in the statistical literature by using hierarchical models that accommodate truncated or censored observations. [4] propose a bias correction method based on a compound Poisson model for count data that includes an area-specific reporting probability, whose uncertainty is accounted for in the model. In the proposal, areas are clustered according to their data quality. Since underreporting probabilities might reflect socio-economic, political and/or demographic characteristics of each region, we focus on modelling the underreporting rates in different areas. Building on the idea in [4], we consider a Bayesian semi-parametric extension of the compound Poisson model: following the approach described in [8], we propose a new method to introduce covariates to define the clustering structure for the underreporting probability ε_i.

2 Modelling underreporting probability

The bias problem induced by defectively reported data is widely discussed in the statistical literature: a great variety of models that accommodate truncated or censored

observations has been developed. Several extensions of the Poisson model have been proposed for counts subject to underreporting ([1], [2], [7]). The compound Poisson model (CPM) is an alternative approach to deal with potentially underreported counts. It allows for the joint modeling of the event occurrence rates and the associated reporting probabilities. [4] discussed issues about the identifiability of the CPM and proposed an alternative Bayesian hierarchical formulation of the CPM that only requires reliable prior information about the reporting process for areas experiencing the best data quality. They rely on the fact that areas within the region under study can be clustered into homogeneous groups reflecting the data quality of each area. The clusters may be defined based on experts' opinion or applying some clustering technique to auxiliary data quality indicators provided by previous studies and surveys.

2.1 A compound Poisson model for underreporting probabilities

Consider a region consisting of m areas and denote by Y_i the observed counts in area i ($i = 1, ..., m$). Denote by T_i the number of events of interest, and assume

$$Y_i \sim Poisson(E_i\theta_i),$$

where θ_i is the relative risk and E_i is a known offset representing the expected number of events in the $i-$th area. T_i is possibily underreported, in the sense that for each record $j = 1, \ldots, T_i$ in area i the event is actually reported with probability ε_i, namely $Z_{ij} \sim Bernoulli(\varepsilon_i), i = 1 \ldots, T_i$ independently, and independent on T_i.

The observed counts are defined as $Y_i = \sum_{j=1}^{T_i} Z_{ij}$ and therefore, $Y_i|T_i, \varepsilon_i \sim Bin(T_i, \varepsilon_i)$. Combining with 2.1 and marginalising over T_i this is the so called compound Poisson model (CPM), under which

$$Y_i|\theta_i\varepsilon_i \sim Poisson(E_i\theta_i\varepsilon_i).$$

Let us further assume that we can relate the relative risks to a set of covariates $X_1, ..., X_p$:

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}.$$

As described above, the parameter ε_i defines the reporting probability in the $i-$th area: low values of ε_i indicate areas whose observed counts are underreported. [4] assume that areas can be a-priori clustered according to their data quality. Indeed, they specify

$$\varepsilon_i = 1 - h_i^T \gamma$$

where $(h_{1i}, ..., h_{Ki})^T$ is the cluster indicator defined according to the following split-coding scheme: if area i belongs to cluster j then $h_{li} = 1$ for all $l \leq j$ and $h_{li} = 0$ otherwise. They also assume that $\gamma \in [0, 1)$ and $\sum_{j=1}^{K} \gamma_j < 1$ to ensure non-null mean for the associated Poisson distribution.

To ensure identifiability of the model, informative priors about the reporting probabilities and the induced clustering structure among the areas are specified: in particular, [4] fix the number of cluster and model the probability of underreporting according to a-priori knowledge about data quality. Strongly informative priors are defined, especially for the areas that are supposed to be characterized by the best data quality.
However, as revealed by the real data application in [4], parameter estimates do sensibly vary according to the number of clusters and to cluster assignment. Besides this, it might be the case that prior information about data quality is not available. In this work, building on the same compound Poisson model, we consider an alternative approach: we specify a clustering structure for the reporting probability ε_i following a non parametric approach based on a dependent Dirichlet process, that allow the aggregating property of the DP to depend on covariates. Although such a specification is more complex from a theoretical as well as computational point of view, it significantly increases the flexibility of the model since it does not require a-priori knowledge of the number of clusters and it defines the clustering structure in a complete nonparametric, data driven way. Indeed, the clustering is induced by introducing covariates in the stick breaking construction of the Dirichlet process.

3 The proposed model

Let $\varepsilon^n = (\varepsilon_1, ..., \varepsilon_m)$ and $Z^n = (z_1, ..., z_m)$ denote, respectively, the entire vector of the reporting probabilities and the covariate Z used as predictor for ε.

A simple nonparametric model can be defined by introducing a DP model on the reporting probabilities:

$$Y_i|\theta_i\varepsilon_i \sim Poisson(E_i\theta_i\varepsilon_i).log(\theta_i) = \beta_i + \beta_1 X_{1i} + ... + \beta_p X_{pi} \quad (1)$$

$$\varepsilon_i|G \sim iid\ G \quad (2)$$

$$G|\alpha, G_0 \sim DP(\alpha, G_0) \quad (3)$$

For any measurable set B, the DP process has the well known stick-breaking representation [9]

$$G(B) = \sum_{j=1}^{\infty} w_j \delta_{\eta_j}(B)$$

where $\delta_{\eta_j}(\cdot)$ is the Dirac measure at η_j and $w_j = V_j \prod_{l<j}[1 - V_l]$ with $V_j|\alpha \overset{i.i.d.}{\sim} Beta(1, \alpha)$ that implies an infinite mixture representation of the distribution of the Y_i.

[3] propose a modification of the well known stick-breaking representation of the DP in which the weights are made dependent on covariates; this is achieved replacing the Beta random variables by normally distributed random variables transformed through the normal cdf. The resulting measure is defined as the probit-stick breaking (PSB) process, see also [9]. As described by [8], [3] allow for dependence on

covariates via the introduction of independent Gaussian processes indexed by the covariates as specified in the following formula:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left\{ \Phi(\eta_j(z)) \prod_{l<j} [1 - \Phi(\eta_l(z))] \right\} \delta_{\varepsilon_j}(\cdot)$$

where $\eta_j(z) = z'\gamma$.
As already discussed in [4], one of the main problems of the CPM is the lack of identifiability. Two main approaches can be used to obtain an identifiable model when sets of covariates, X and H, are used to model the relative risk θ and the reporting probability ε. The first one requires extra information desumed from independent validation datasets and requires non-overlapping and incorrelated X and H. The second one requires informative prior distribution for the overall mean reporting rate which is sufficient to complete the identifiability conditions. In our approach, identifiability is achieved by using two disjoint sets of covariates for estimating the risk and the reporting probability. Moreover, slight informative prior about the intercept parameter β_0 is specified ($\beta_0 \sim N(0, 10)$). Sensitivity analysis with respect to this choice reveal that the model is robust with respect prior specification.

4 Data application and results

We apply our model to data analyzed in [4]: our goal is to accurately estimate the relative risk of early neonatal mortality (ENM) in Minas Gerais State (MG), Brazil, one of the more deprived area of the country. Experts believe that the quality of infant mortality information produced in this country is usually underreported, especially in some areas located in northern and south-Eastern regions of the state. The observed counts of early neonatal deaths have been collected for the $m = 75$ areas from 1999-2001 and 2009-2011. In this work we consider data from 1999-2001. The ENM relative risk assumes a log-linear regression structure which includes local and spatial random effects. Five covariates are introduced in this regression model: the Municipal Human Development Index (MHDI), the proportion of mothers with more than twelve years of formal education (MomEduc), the proportion of children with weight at birth smaller than 2.5 Kg (LowWeight), the proportion of children who were born with some congenital anomaly (Anomaly) and the proportion of mothers who made seven or more prenatal visits during the pregnancy (Prenatal). In [4] the clustering structure is defined according to the quantiles of an adequacy index (AI) introduced by [5] as a measure of the quality of infant mortality data collected in Minas Gerais. In our approach this variable is used in the Dirichlet process for defining both the number of clusters as well as the clustering structure. Informative prior about the intercept parameter is specified according to the obtained estimates with a simple Poisson model.
Table 1 shows the posterior means of the parameters and the corresponding 95% credible intervals. As expected, the two models agree on the fact that only the co-

variate MHDI shows to be significant (likely non-zero effect) to explain the ENM risk for the period 1999–2001. The effect of the covariate MDHI is negative in both periods, indicating that the highest the MHDI, the smallest the ENM risk. According to the approach in [4], the model with $K = 8$ clusters resulted to be the best fitting one while the proposed model selects $K = 5$ clusters, showing a WAIC smaller than the one of the competing model. Also, risk estimates are substantially overlapping confirming that the ENM risks in the poorest areas (North and Northeast) are higher than the ones obtained for more developed regions of Minas Gerais.
These results reveal that the proposed approach is accurate and particularly suitable when there are no prior information about data quality.

Table 1 Parameter estimates and 95% credible intervals (in brackes) for the model proposed in [4] and the proposed model.

Parameter	Lopes et al. model $K = 8$ $WAIC = 603.091$	Proposed model $K = 5$ $WAIC = 598.360$
Intercept	1.986 (1.624; 2.348)	0.852 (0.092; 1.612)
MHDI	-3.369 (-3.991; -2.747)	-1.635 (-2.868; -0.402)
MomEduc	-0.033 (-1.263; 1.197)	1.452 (-4.469; 1.317)
LowWeight	-0.095 (-1.279; 1.089)	2.206 (-3.932; 6.836)
Anomaly	2.450 (-11.972; 16.872)	2.207 (-15.644; 20.057)
Prenatal	0.104 (-0.230; 0.528)	0.507 (-0.0311; 1.044)

References

1. Bailey, T. C., Carvalho, M. S., Lapa, T. M., Souza, W. V., and Brewer, M. J. Modeling of under-detection of cases in disease surveillance. Annals of Epidemiology **15(5)**, 335–343, (2005).
2. Caudill, B. S. and Mixon Jr., F. G. Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. Empirical Economics **20(2)**, 183–196, (1995).
3. Chung, Y. and Dunson, D. B. Nonparametric Bayes conditional distribution modeling with variable selection. Journal of the American Statistical Association **104**,1646–1660, (2009).
4. de Oliverira, G.L., Argiento, R., Loschi, R.H. Bias correction in clustered underreported data, Bayesian Analysis TBA, 1–32, (2020)
5. Franca, E., Abreu, D., Campos, D. and Rausch, M. C. Avaliacao da Qualidade da informacao sobre a mortalidade infantil em Minas Gerais: Utilizacao de uma metodologia simplificada (available in Portuguese). Revista Medica de Minas Gerais **16(1 suppl. 2)**, 15 –22, ((2006).
6. Hart,J.T. The inverse care law, Lancet **7696**, 405–412, (1971).
7. Oliveira, G. L., Loschi, R. H., and Assuncao, R. M. A random-censoring Poisson model for underreported data. Statistics in Medicine **36(30)**, 4873–4892, (2017).
8. Quintana, F., Mueller, P., Jara, A., MacEachern, S. The dependent Dirichlet process and related models. arXiv:2007.06129, (2021)
9. Sethuraman, J.A constructive definition of Dirichlet prior. Statistica Sinica **2**, 639–650, (1994).