Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

A Short Demo Article: Regression Models for Count Data in R

Juan-Ramón González
ISGlobal

Dolors Pelegrí
ISGlobal

Isaac Subirana
CIBERESP

Abstract

.. to be completed ..

Keywords: JSS, style guide, comma-separated, not capitalized, R.

1. Introduction

2. Methods

3. Real data example

... to be completed ...

4. Simulation studies

4.1. Simulation methods

In order to assess the validation of the proposed method, two simulation studies under different scenarios have been performed. Results are compared to two other alternative approaches consisting on (1) IMPUTE: imputing the average of available values to missing individuals for each variable, and (2) COMMON: analyse the individuals with complete data in all data sets. For both, data has been generated similar to as described in Velden and Bijmolt (2006). Detailed steps are listed below.

Performance of each method was assessed comparing its results applied to data with missing rows to the results obtained if the full data were available (case 1). In addition, discrimination capacity of each method was evaluated in another simulation study (case 2) where two groups were generated.

Case 1

- **Step 1:** generate a matrix,Y_{latent}, of n by 2 independent variables from a standarditzed normal distribution, which are the two latent variables.
- **Step 2:** generate two matrices, b_1 and b_2 with dimensions 2 by p_1 and 2 by p_2 of coefficients ranging from -1 to 1 under a uniform distribution.
- **Step 3:** Compute X_1 and X_2, of dimensions n by p_1 and n by p_2, respectively, pos-multiplying Y_{latent} by coefficient matrices b_1 and b_2.
- **Step 4:** Add noise to X_1 and X_2 by adding a gnerated normal value of zero mean and σ_2 standard deviation. At this point full data is obtained.
- **Step 4:** Randomly select a proportion of rows for X_1 and X_2 (not the same rows) to be declared as missing individuals.

Data were generated under the following scenarios:

- Fixing the number of individuals n to 500.
- Varying the number of variables p,q to 50 and 100.
- Varying the noise σ to 0.125 and 0.250.
- Varying the proportion of missing individuals p to 0.1, 0.2 and 0.3.

So, 2x2x3 = 12 escenarios were simulated, and for each simulated scenarios 300 data sets were generated.

Generalized Canonical Correlation Analyses was performed for each generated data and canonical latent variables were computed, using full data ("FULL"). Then, the three methods (MGCCA, IMPUTE, COMMON) were applied to the data with missing rows. Finally, performance of each of the three methdos were measured in terms of how diffent are the scores of estimated latent canonical variables with the ones obtained with the full data as follows:

$$dist = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_{full}[i1] - \hat{y}_{approach}[i1]\right)^2 + \left(\hat{y}_{full}[i2] - \hat{y}_{approach}[i2]\right)^2$$

where "approach" referes to MGCCA (proposed method), IMPUTE or COMMON.

Note that when computing "dist" for COMMON approach, "n" is the number of individuals with complete data, so rows of $\hat{y}_{full}$ and $\hat{y}_{COMMON}$ corresponds to these individuals.

Case 2

Another generation data process similar to the one described above was performed but now two groups are distinguished, and methods are assessed in terms of discriminate these groups.

All steps are the same except the step 1, where first $n/2$ rows (individuals) of Y_{latent} matrix are generated under a normal distribution with mean equals to $\delta/2$ and the rest of the rows under a normal distribution with mean equals to $-\delta/2$.

Once the data is generated, a MANOVA anlyses is performed to test difference in means of estimated canonical variables scores among the two groups.

In these simulation study, number of generated individuals and variables were fixed to $n = 500$, and $p = q = 50$, respectively. Noise standard deviation was fixed to $\sigma = 0.2$. While the varying parameters where the difference in means of groups $\delta = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and proportion of missing individuals $p = \{0.1, 0.3\}$

4.2. Simulation results

Case 1

.. to be completed ..

Case 2

.. to be completed ..

5. Summary and discussion

... to be completed ...

Computational details

... specify R version and used packages ...

... consumed time for simulation. ? analyses of real data ?

Acknowledgments

References

Velden M, Bijmolt T (2006). "Generalized canonical correlation analysis of matrices with missing rows: a simulation study." *Psychometrika*, **71**(2), 323–331. doi:10.1007/s11336-004-1168-9.

A. More technical details

... not sure if necessary to include an appendix...

Affiliation:

Juan-Ramón González
ISGlobal
and
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://eeecon.uibk.ac.at/~zeileis/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd