plsRbeta, PLS Beta Regressions in *R*

Frédéric Bertrand[1*] , Myriam Maumy-Bertrand[1]

1. Université de Strasbourg & CNRS
*Contact author: frederic.bertrand@math.unistra.fr

Keywords: Partial least squares regression, beta regression models, bootstrap, high dimensional data, R language package

Many responses, for instance experimental results or economic indices, can be naturally expressed as rates or proportions whose values must hence lie between zero and one. The Beta regression often allows to model these data accurately since, according to Johnson et al. [3] : "Beta distributions are very versatile and a variety of uncertanties can be usefully modelled by them. This flexibility encourages its empirical use in a wide range of applications".
Several recent articles focused on Beta regression : see Ferrari and Cribari-Neto [2] for an introduction and Kosmidis and Firth [4] and Simas et al. [5] for more insights on estimation techniques. Yet, as any of the usual regression model, it cannot be applied safely in case of multicollinearity and not at all when the model matrix is rectangular with more variables than subjects. These situations are frequently found from chemistry to medicine through economics or marketing. To circumvent this diffculty, we derived an extension of PLS regression [7] to Beta regression models, Bertrand et al. [1]. Two non linear extensions of PLS Beta regression are introduced using kernel techniques following ideas that were successfully applied in Tenenhaus et al. [6] to PLS logistic regression. PLS Beta Regression, as well as several other tools, such as cross validation, bootstrap or kernel techniques, is available for the *R* language in the **plsRbeta** package.

We will provide a simulation study and an application to an allelotyping dataset collected on 93 subjects suffering from various types of lung cancer. Since the response is the tumoral cellularity, which is a rate, and the dataset is rectangular, a PLS Beta regression model is relevant.

References

[1] Bertrand, F., Meyer, N., Beau-Faller, M., El Bayed, K., Namer, I.-J. and Maumy-Bertrand, M. (2012). *Journal de la Société Française de Statistique*, Régression Bêta PLS 154(3), 143–159.

[2] Ferrari, S.L.P. and Cribari-Neto, F. (2004). *Journal of Applied Statistics*, 31(7), 799.

[3] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, vol. 2, 2nd ed. New York, Wiley.

[4] Kosmidis, I. and Firth. D. (2010). *Electronic Journal of Statistics*, 4, 1097, (2010).

[5] Simas, A.B., Barreto-Souza, W. and Rocha, A.V. (2010). *Computational Statistics & Data Analysis*, 54(2), 348.

[6] Tenenhaus, A., Giron, A., Viennet, E., Béra, M., Saporta, G. and Fertil, B. (2007). *Computational Statistics & Data Analysis*, 51, 4083.

[7] Wold, S., Sjöström, M. and Eriksson, L. (2001). *Chemometrics and Intelligent Laboratory Systems*, 58, 109.