Dealing with collinearity in quantile regression

Domenico Vistocco

domenico.vistocco@unina.it

Abstract

The talk aims to explore the collinearity problem in quantile regression. Collinearity among predictors is one of the main problems associated with multiple linear regression. Its effect on estimates, standard errors, computational accuracy, fitted values and predictions is well investigated in classical regression [6]. This is not the case for quantile regression (QR) [8, 7, 3, 5], where the contributions in the literature essentially focused on variants of ridge regression [2], or on the use of variable selection techniques [1]. We face the problem from a different perspective trying to preserve the whole set of variables eliminating any redundancy that is unnecessary and detrimental to the estimation process through principal component regression (PCR). The basic idea of PCR is to find some linear combinations of the original variables and use them as regressors to predict the response variable. The identification of the main components takes place on the basis of dimensionality reduction techniques. PCR consists of two steps: a PCA applied to the predictor matrix, and a subsequent regression of the response variable on the first principal components, those that explain most of the variability [9]. Therefore, the extension of PCR to the context of QR [4] is straightforward: the extraction of the main components from the predictor matrix occurs in the same way, while QR is used for the regression of the response variable on the extracted components. This technique also offers the side-effect of easy to interpret graphical representations of the results. In particular, the loading plot makes it possible to represent regression coefficients on the principal components, so to understand which are the most critical variables in constructing the principal components and, therefore, in predicting the response variable. The score plot is instead used to depicts the coordinates of the observations on the principal components, with the aim to detect similarities and differences between the different statistical units. The score plot is also useful for outlier detection. The talk explore the effect of collinearity on standard errors, as well as in terms of in-sample and out-of-sample prediction ability, and of estimate bias, through a simulation study and a case study showing how the set of starting predictors, highly correlated with each other, can be suitably synthesized into new variables, representing latent dimensions of observed data, that can be effectively used to predict the response.

References

[1] Alhamzawi, R. and Yu, K. [2012], 'Variable selection in quantile regression via Gibbs sampling', *Journal of Applied Statistics* **39**(4), 799–813.

[2] Bager, A. [2018], 'Ridge parameter in quantile regression models: An application in biostatistics', *International Journal of Statistics and Applications* **8**(2), 72–78.

[3] Davino, C., Furno, M. and Vistocco, D. [2013], *Quantile regression: theory and applications*, Vol. 988, John Wiley & Sons.

[4] Davino, C., Romano, R. and Vistocco, D. [2022], 'Handling multicollinearity in quantile regression through the use of principal component regression', *Metron* **80**(2), 153–174.

[5] Furno, M. and Vistocco, D. [2018], *Quantile regression: estimation and simulation, Volume 2*, Vol. 216, John Wiley & Sons.

[6] Gunst, R. F. [1983], 'Regresion analysis with multicollinear predictor variables: definition, derection, and effects', *Communications in Statistics-Theory and Methods* **12**(19), 2217–2260.

[7] Koenker, R. [2005], *Quantile regression*, Vol. 38, Cambridge University Press.

[8] Koenker, R. and Bassett Jr, G. [1978], 'Regression quantiles', *Econometrica: Journal of the Econometric Society* pp. 33–50.

[9] Næs, T. and Mevik, B.-H. [2001], 'Understanding the collinearity problem in regression and discriminant analysis', *Journal of Chemometrics: A Journal of the Chemometrics Society* **15**(4), 413–426.