A COMPARISON AMONG ESTIMATORS FOR LINEAR REGRESSION METHODS

Marilena Furno[1] and Domenico Vistocco[2]

[1] Dipartimento di Agraria, Università degli Studi di Napoli – Federico II, (e-mail: marfurno@unina.it)

[2] Dipartimento di Economia e Giurisprudenza,Università degli Studi di Cassino e del Lazio Meridionale, (e-mail: vistocco@unicas.it)

KEYWORDS: Quantile regression, Expectile regression, M–estimators, M–quantile.

1 Estimators for linear regression methods

Koenker & Basset, 1978 introduce the quantile regression estimator, that allows to have a more complete view of the effects a set explicative variables exerts on the response, not only on average but at different points of the conditional distribution: the conditional quantiles. The core of quantile regression is the use of an asymmetric check function that moves the regression line above or below the conditional median, allowing to consider location, scale and shape effects in the study of a statistical relationship. Quantile regression is increasingly implemented due to the variety of its possible applications, as evidenced by the growing number of related papers in recent years. For further details see Koenker, 2005, Hao & Naiman, 2007 and Davino *et al.* , 2013.

A computationally simpler alternative to quantile regression, introduced by Newey & Powell, 1987, is provided by the expectile regression. Expectiles allow the analysis of a regression model at various points of the conditional distribution through the introduction of an asymmetric weighting system. Analogously to the asymmetric check function of the quantile regression estimator, such weighting system moves the least squares regression line, as estimated by ordinary least squares, above or below the regression passing through the conditional average. Compared to quantile regression, the expectile regression is computationally convenient and it still allows to characterize the complete conditional distribution of the response. The main characteristic of expectiles is the adoption of the L_2 norm and this causes the lack of robustness of the expectile with respect to the quantile regression estimator.

The class of robust regression estimators, the M-estimators (Huber, 1981), computes the regression at the conditional mean meanwhile curbing the impact of outliers on the estimated coefficients. Once again this estimator considers a weighting system to detect and bound outliers while estimating the regression coefficients. Breckling & Chambers, 1988 propose to merge the M–estimators and the expectile approach. Even if both methods are implemented within the least squares framework, robustness is ensured by the introduction of a weighting system to control the outlying observations. The asymmetric weighting system of expectiles is combined with weights bounding outliers and this allows to compute a robust regression away from the conditional mean.

Along with the above estimators, it is worth mentioning the modal linear regression (Kemp & Santos Silva, 2012, Yao & Li, 2014). Here the focus is on modeling the conditional mode of the response variable, and it is well adapt in situations where conditional distributions are highly skewed: exploiting the mode features, modal regression reveals robust to outliers, in particular to heavy-tailed conditional error distributions.

2 The loss functions

Quantile and expectile regression can be interpreted in a common framework using the following loss function:

$$\min \sum_{i=1}^{n} w_\theta |y_i - \beta(\theta) x_i|^r = \sum_{i=1}^{n} w_\theta |e_i|^r$$

where $r = 1$ in case of quantile regression and $r = 2$ in case of expectile regression. For both the estimators, the asymmetric weights are defined as follows:

$$w_\theta = \begin{cases} 1-\theta, & \text{if } e_i < 0 \\ 0, & \text{if } e_i = 0 \\ \theta, & \text{if } e_i > 0 \end{cases}$$

The regression line is estimated at the θ quantile or expectile. For instance, when $\theta = 0.75$, positive errors have a larger weight and the regression line is attracted toward the upper tail of the error distribution. The role and the interpretation of θ in the above formula is naturally different in the two type of regression models, referring to the conditional quantile of the response variable for quantile regression and interpretable as a measure of location on the response conditional distribution for expectile regression.

The quantile regression problem can be solved through linear programming methods or interior point methods (Koenker, 2005); iteratively weighted least squares is instead used to compute the expectiles (Newey & Powell, 1987).

As mentioned, the use of a squared loss function involves a lack of robustness of expectiles as anomalous values exert a strong impact on the objective function. Huber, 1981, with the class of M–estimators, introduces robustness in the ordinary least squares estimator by bounding large residuals, thus controlling the impact of outliers. The M–estimators are commonly computed iteratively through weighted least squares.

M–quantiles merge together the M–estimators and the expectiles approach, allowing to compute robust regressions passing through different points of the conditional distribution. The use of an asymmetric weighting system allows to preserve robustness by controlling outlying observations and to move the regression line beyond the conditional mean. In detail, the objective function of the M–quantile estimator computed at θ is defined as:

$$\sum_{i=1}^{n} \rho(u_i)\,[(1-\theta)I(u_i \leq 0) + \theta I(u_i > 0)]$$

where $\rho(.)$ curbs the outliers and $I(.)$ is the indicator function. The first derivative of the M–quantile estimator is:

$$\sum_{i=1}^{n} \psi_\theta(u_i)x_i = 0$$

with:

$$\psi_\theta(u_i) = \begin{cases} (1-\theta)\,\psi(u_i), & \text{if } u_i \leq 0 \\ \theta\,\psi(u_i), & \text{elsewhere} \end{cases}$$

In the last equation u_i are the standardized residuals $\frac{e_i}{\sigma}$, with σ computed by $MAD = \frac{median|e_i|}{0.6745}$ and $\psi(u_i)$ defines the Huber M–estimator or any other of the functions belonging to the class of M–estimators.

3 Simulations

A Monte Carlo study compares the quantile regression, the expectile regression and the M–quantile estimators in case of heterogeneous error models. Considering the linear regression model, the three estimators are compared in case of non–normal errors and in presence of heteroscedasticity. In particular different degrees and types of skewness, along with different heteroschedasticity patterns are taken into account in order to show if and how the M–quantile estimator is able to stand in case of heterogeneous data.

The results are not shown in this short version of the manuscript for lack of space.

References

BRECKLING, J., & CHAMBERS, R. 1988. M–quantiles. *Biometrika*, **75**, 761–71.

DAVINO, C., FURNO, M., & VISTOCCO, D. 2013. *Quantile regression: theory and applications*. Wiley Series in Probability and Statistics. Wiley.

HAO, L., & NAIMAN, D.Q. 2007. *Quantile regression*. Sage Pubblications, Inc.

HUBER, P. 1981. *Robust Statistics*. Wiley.

KEMP, G., & SANTOS SILVA, J. 2012. Regression towards the mode. *Journal of Econometrics*, **170**, 92–101.

KOENKER, R. 2005. *Quantile regression*. Cambridge University Press.

KOENKER, R., & BASSET, G. 1978. Regression quantiles. *Econometrica*, **46**(1), 33–50.

MOSTELLER, F., & TUKEY, J. 1977. *Data analysis and regression: a second course in Statistics*. Reading, MA: Addison–Wesley.

NEWEY, W., & POWELL, J. 1987. Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819–47.

YAO, W., & LI, L. 2014. A New Regression Model: Modal Linear Regression. *Scandinavian Journal of Statistics*, **41**, 656–671.