**Extreme bound analysis based on correlation coefficient for optimal regression model**

**Abstract**

**1. Introduction**

**2. Methodology**

I propose two concepts of correlation such as local correlation and global correlation. Local correlation is also called model correlation, which implies fitness of a target regressive parameter with subject to a given regression model. Note, regressive parameter ***α*** = (*α*0, *α*1, *α*2,…, *αn*) is the set of regressive coefficients corresponding to regressors ***x*** = (*x*1, *x*2,…, *xn*) and let *z* and be the responsor and its estimate, respectively. Given regression model *k*, let *Rk*(*xj*, ) and *Rk*(, *z*) be the correlation between *xj* and and the correlation between and *z* within model *k*, respectively.

Suppose the estimate of the *j*th coefficient *αj* with regard to regressor *xj* is , let *Rk*(*xj*, z) be the local correlation of *xj* and *z* within model *k*. Obviously, *Rk*(*xj*, *z*) reflects fitness or appropriateness of the regressive coefficient estimate regarding model *k*. The local correlation *Rk*(*xj*, *z*) is defined as product of *Rk*(*xj*, ) and *Rk*(, *z*) as follows:

Indeed, local correlation is a conditional correlation of a regressor along its estimated coefficient given the condition which is the estimated regression model and so, the intermediate variable representing such condition is the estimated response . For *K* models which are estimated, averaged local correlation is calculated as follows:

Global correlation implies fitness of the target regressive parameter without concerning any regression models. Let *R*(*xj*, *z*) denote the global correlation between regressor *xj* and responsor *z*, which is defined as usual correlation coefficient as follows:

A regressor *xj* along with its implicit regressive coefficient *αj* are good if they can give sufficient explanatories to possible models and they can be more independent to reflect the responsor *z*. In other words, the first condition of sufficient explanatories to possible models is represented by the local correlation and the second condition of independent reflection is represented by the global correlation. Therefore, the fitness of *xj* and *αj* are defined as product of the averaged local correlation and the global correlation *R*(*xj*, *z*) follows:

The larger the fitness *φj* is, the better the implicit estimate is, and the better the regressor *xj* is. Good regressors *xj* (also *αj* or ) which have large enough fitness values *φj* are called robust regressors. Consequently, Regressive Expectation Maximization with RObust regressors (REMRO) algorithm searches for robust regressors and sorts them according to descending ordering with their fitness values *φj* as searching criterion. Another problem is how to produce *K* models to calculate the averaged local correlation . Fortunately, Sala-i-Martin generated a set of *K* combinations of doubtful regressors which need to be checked their fitness. Each model in *K* models is estimated with each combination of doubtful ones and estimation method can be least squares method as usual. Moreover, REMRO can resist incomplete data because it applies REM algorithm into filling missing values for both regressors and responsor by estimated values based on ideology of expectation maximization (EM) algorithm. Let free set ***D*** be the set of regressors which is compulsorily included in the regression model and let focus set ***F*** = ***X***\ ***D*** be the complement of ***D*** with subject to the entire set ***X***. Let *c* = |***C***| be the number of regressors in each combination set ***C*** taken from doubtful set ***B*** = ***F*** \ {*xj*} where *xj* is current focused regressor, the following is flow chart of REMRO algorithm.

**Figure 2.1.** Flow chart of REMRO

Indeed, REMRO estimates fitness values of focused regressors in ***F*** and then builds up regression model with high fitness regressors. The final regression model estimated by REMRO with only robust regressors is called optimal regression model. Each combination suggested in literature includes three doubtful regressors, *c* = 3. Here I suggest the size *c* of each combination is half the cardinality of focus set ***F*** and hence, the number of models is determined as follows:

Note, the notation represents lower integer of given real number.

Sala-i-Martin estimated the fitness of estimate as the value of cumulative density function of *αj* at 0, denoted as cdf(*αj* =0 | ) followed by calculating the mean and the variance of *αj* based on likelihood function over *K* models.

Especially, Sala-i-Martin mentioned the variance as averaged variance of *K* models. When REMRO is tested with Sala-i-Martin method, I here improve Sala-i-Martin formulation by estimating only based on *K* distributed values of because the averaged variance of *K* models does not reflect variation of regressors. For instance, give *K* models and suppose each estimate of *αj* within model *k* is , the variance is calculated as follows:

Where *Lk* is likelihood function of model *k* and the mean is stilled followed Sala-i-Martin formulation.

In practice, *Lk* is replaced by logarithm of likelihood function *lk* = log(*Lk*) in order to prevent producing very small number due to large matrix data with many rows.

REMRO applies REM algorithm into computing regressive estimates = (, ,,…, )*T* and REM, in turn, applies EM algorithm to resist missing values. It is necessary to describe shortly REM here.

**3. Experimental results and discussions**

In this experiment, REMRO is tested with Sala-i-Martin given absolute mean error (MAE) as testing metric. MAE is absolute deviation between original response z in matrix data and estimated response produced from regression model.

**4. Conclusions**

**References**