Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 3 (2016), pp. 3305–3314
© Research India Publications
http://www.ripublication.com/gjpam.htm

Lasso and Ridge Quantile Regression using Cross Validation to Estimate Extreme Rainfall

Hilda Zaikarina, Anik Djuraidah, and Aji Hamim Wigena

Department of Statistics,
Bogor Agricultural University,
Bogor, Indonesia.

Abstract

In quantile regression there should be no multicollinearity in predictor variables. The lasso or ridge regularization can overcome the problem of multicollinearity. An optimum lasso or ridge coefficient can be estimated through cross validation method. However, this method is unstable when the cross validation process is repeated. Percentile method is a method to stabilize cross validation but is not the best method to predict an extreme value. This paper discusses the use of lasso and ridge regularizations in quantile regression with a modified percentile cross validation. The optimum lasso and ridge coefficients are determined based on cross validation error minimum. The results show that the quantile regression with ridge regularization was better than that with lasso regularization to estimate extreme rainfall. The values of RMSEP (Root Mean Square Error of Prediction) of ridge regularization are 16.53, 18.36, and 26.26 at Q(0.75), Q(0.90), and Q(0.95) respectively, while those of lasso regularization are 15.16, 21.19, and 37.24 at the same quantiles.

AMS subject classification:
Keywords: Quantile Regression, Lasso, Ridge, Cross Validation, Percentile.

1. Introduction

Global circulation model (GCM) output such as precipitation which is global scale or low resolution data cannot be used directly to describe the condition in a local area. GCM output can be used as a source of information in statistical downscaling (SD). A model in SD relates functionally the precipitation of GCM output as predictor variables with rainfall data as response variable. The model can be written as multiple regression, $y = f(X)$, with y and X are the rainfall data and GCM precipitation respectively.

The distribution of rainfall data is asymmetric and heavy right tail. To predict extreme rainfall is not the best way using ordinary least square regression that provides a convenient method of estimating such conditional mean model [9]. Quantile regression fits to describe the extreme rainfall because it can explain the nature of rainfall in any quantile. Quantile regression can provide satisfactory results as least square regression on the condition of the entire assumptions are fulfilled [11].

High dimension of precipitation of GCM output leads to multicollinearity [14] that makes the solution of quantile regression becomes not unique. The solution to handle that problem is to reduce the dimension [1], to select variables [2][4][13], and to shrinkage coefficient [5] [6][12]. In recent years, much interest has focused on shrinkage methods such as lasso and ridge methods [11]. A method commonly used is cross validation (CV) to determined lasso and ridge coefficients. The CV is used because of the limited amount of data that can be used in an analysis. One set of data is divided into modeling data and validation data [3]. Simulations carried out by [10] to build generalized linier model with lasso regularization. It was found that the process of CV is not stable in choosing a lasso coefficient when the processes are repeated. Some variation is expected because the grouping of data in the CV process is random [8]. Lasso percentile method was proposed to deal with the instability CV [10] and found that from a hundred replicates the best lasso coefficient to build linear model is over than P(0.75). However, the selection lasso coefficient over than P(0.75) is not the best to predict extreme value in quantile regression. So, this paper concerns with finding the best criteria not only lasso coefficient but also ridge coefficient to predict extreme value in quantile regression model using a modified percentile method. The main focus is on the extreme values at Q(0.75), Q(0.90), and Q(0.95).

2. Methodology

2.1. Quantile Regression

Quantile regression introduced by Koenker and Bassett in 1978 is an extension of the quantile function. Quantile regression built a comprehensive strategy to capture the whole picture regression [9]. Quantile regression coefficients, β_τ, are predicted based on [7]:

$$\hat{\beta}_\tau = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^T \beta_\tau) \tag{2.1}$$

with $i = 1, \ldots, n$ and $\rho_\tau(.)$:

$$\rho_\tau(.) = \begin{cases} (y_i - x_i^T \beta_\tau)(\tau - 1) & if (y_i - x_i^T \beta_\tau) < 0 \\ (y_i - x_i^T \beta_\tau)\tau & if (y_i - x_i^T \beta_\tau) \geq 0 \end{cases} \tag{2.2}$$

Quantile regression with lasso and ridge regularization used lasso and ridge coefficients to build quantile regression model. Solution of lasso coefficient can be written in Lagrangian form as shown below:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta_\tau) + \lambda_{lasso} \sum_{j=1}^{p} |\beta_{\tau,j}| \tag{2.3}$$

and solution of ridge coefficient can be written in Lagrangian form as follows:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta_\tau) + \lambda_{ridge} \sum_{j=1}^{p} (\beta_{\tau,j})^2 \tag{2.4}$$

2.2. Percentile Method

Basically, the percentile method is CV process that repeated in more than once. There are two algorithms of percentile method that proposed by Roberts and Nowak in 2013. Both algorithms are simple modification of standard lasso, called percentile lasso. One of them was applied in this research. The algorithm is as follows [10]:

1. For $m = 1$ to M do
    - Randomly assign observations to folds for cross validation.
    - Fit the standard-lasso using this fold assignment.
2. Let $\hat{\lambda}_m$ denote the optimal tuning parameter (lasso coeficient) obtained from standard-lasso
3. Let $\Lambda = (\hat{\lambda}_1, \hat{\lambda}_1 00)$ denote the 100 values of $\hat{\lambda}_m$ with CVE for each $\hat{\lambda}$
4. Compute $\lambda(\theta)$, the θ-percentile of $\Lambda(M)$
5. The percentile lasso solution is the solution of the standard-lasso fitted with $\lambda = \lambda(\theta)$.

Roberts and Nowak proposed to choose $\theta \geq 0.75$. In prediction extreme rainfall, $\theta \geq 0.75$ is not optimal because contain a largest root mean square error prediction (RMSEP). Beside that, it is difficult to choose fixed Î̂¸ that has minimum of RMSEP. It make modifications carried out on point (3). Each $\hat{\lambda}$ in $\Lambda(M)$ have a cross validation error (CVE). The modifications are choose $\hat{\lambda}$ that has a minimum CVE in $\Lambda(M)$. In this research, percentile method not only applied on standard lasso but also on ridge. We assign value of M as 100. This computation used software R with âŁœhqregâŁž packages. The modification of algorithms as shown below:

1. For $m = 1$ to 100 do
    - Randomly assign observations to folds for cross validation.

Table 1: Models with Data are Built.

Model	Modeling Data (Year)	Actual Quantil Data (Year)
M1	1981-2009	1981-2009
M2	1981-2010	1981-2010
M3	1981-2011	1981-2011
M4	1981-2012	1981-2012

- Fit the standard-lasso or ridge using this fold assignment.

2. Let $\hat{\lambda}_m$ denote the optimal lasso coeficient or ridge coeficient obtained from standard-lasso or ridge.

3. Let $\Lambda = (\hat{\lambda}_1, \hat{\lambda}_{100})$ denote the 100 values of $\hat{\lambda}_m$ with CVE for each $\hat{\lambda}$

4. Solution for lasso or ridge coeficient is λ in Λ that has a minimum of CVE.

3. Data

Data used in this research are local monthly rainfall and precipitation of GCM data. Local monthly data in 1981-2013 at Indramayu, Indonesia, is the average of four weather stations Krangkeng, Sukadana, Karangkendal, and Gegesik. Precipitation of GCM data are consists of monthly rainfall data Climate Model Intercomparison Project (CMIP5) issued by the Dutch KNMI, from the website http://www.climatexp.knmi.nl/ in 1981–2013 with the region's position 18.75°–1.25° South Latitude and 101.25°–116.25°East Longitude. The observed area is a square shaped area of 8-8 grid, which resulting in 64 predictor variables.

There are four models are built to know the consistency quantile regression model. Each model devided rainfall and precipitation of GCM data to modeling data. Actual quantile data are built from rainfall data that grouped by month in Q(0.75), Q(0.9), and Q(0.95). Actual quantile data will compared with predictied of quantile regression model to measure RMSE and RMSEP. The models are shown below:

4. Results

4.1. Modeling

Lasso and ridge quantile regression models established by lasso and ridge coefficients. These coefficients are selected based on modified percentile method for each quantiles and each models. In order to know which regularization gives more influence to model. In Table 2 shows that for each quantile the optimum lasso coefficients are not very much different and that of ridge are also not different. These lasso coefficients are greater than those of ridge coefficients. This means lasso regularization gives more influence

Table 2: List of $\hat{\lambda}$ based on Modified Percentile Method

Quantile	Model	Lasso		Ridge	
		λ_{lasso}	CVE	λ_{ridge}	CVE
Q(0.75)	M1	2.7410^{-2}	21.78	1.1010^{-3}	22.23
	M2	3.1210^{-2}	22.15	1.1110^{-2}	22.51
	M3	3.3110^{-2}	22.13	1.2210^{-2}	22.61
	M4	2.8810^{-2}	22.09	1.1210^{-2}	22.56
Q(0.9)	M1	2.6210^{-2}	12.78	1.8310^{-3}	12.97
	M2	1.9210^{-2}	12.92	2.310^{-3}	13.15
	M3	2.3310^{-2}	12.91	3.510^{-3}	13.18
	M4	2.0310^{-2}	13.01	3.410^{-3}	13.18
Q(0.95)	M1	2.510^{-3}	7.52	310^{-4}	7.61
	M2	1.610^{-3}	7.68	610^{-4}	7.82
	M3	2.610^{-3}	7.72	310^{-4}	7.82
	M4	1.610^{-3}	7.84	1.210^{-3}	7.83

on quantile regression model than ridge regularization. Both lasso or ridge coefficents shows that the greater quantile give the smallest influence.

The models are then used to predict extreme rainfall. The predicted rainfall are compare to the actual quantile data. Figure 1 shows the comparison of RMSE values of all models. Lasso and ridge regularization in Q(0.75) had consistently the smallest RMSE. The highest RMSE occured in Q(0.95) on every model. The values of RMSE that less than 40 for all models show that quantile regression models with lasso and ridge regularization are consistently good. However, quantile regression models with ridge regularization were better than that with lasso regularization.

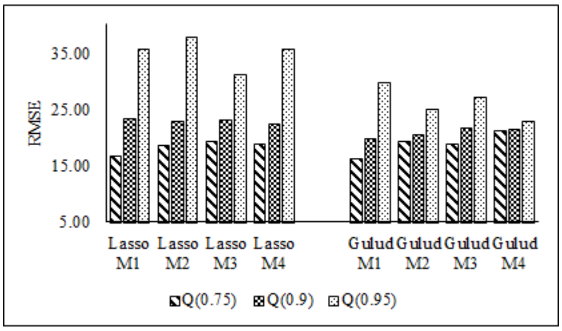

Figure 1: RMSE of Lasso and Ridge Quantile Regression Models

4.2. Extreme Rainfall Prediction

A hundred of lasso and ridge coefficients from Î› used for built quantile regression model to predict extreme rainfall. Model 1 (M1) is used to predict rainfall in 2010, M2 in 2011, M3 in 2012, and M4 in 2013. Predicted extreme rainfall from each models are compared with actual quantile data to get RMSEP, so we have a hundred RMSEP. Lasso and ridge coefficients from modified percentile method is better than percentile method based on RMSEP. Table 3 and Table 4 shows that coefficients from modified percentile method at each quantile are between the minimum and maximum RMSEP, beside that it is never reach maximum RMSEP. This conditions are different with coefficients from percentile method that contain the maximum RMSEP. Although minimum percentile method reach minimum RMSEP in some quantile, it is difficult to draw conclusions the best -percentile for the next prediction because minimum and maximum of RMSEP not consistently in the same percentile. This shows the modified percentile method is good enough to determine optimal lasso and ridge coefficients in establishing a quantile regression model. Figure 2 shows RMSEP and correlation between the predicted and actual quantile data for each model. The results shows RMSEP of each quantile in each models are not very much different. The models with RMSEP smaller than 50 with correlation more than 0.95 can predict extreme rainfall for each quantile accurately. Quantile regression models with ridge regularization are better than those lasso regularization because of the smallest RMSEP and the highest correlation.

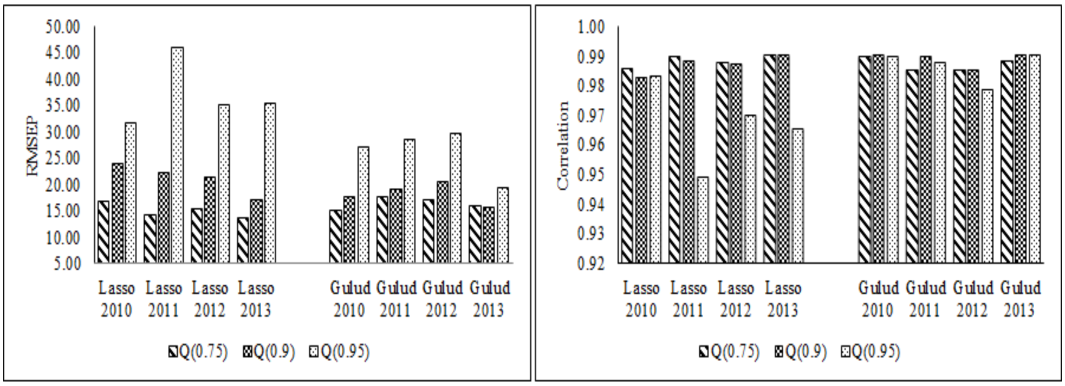

Figure 2: RMSEP and Correlation of Lasso and Ridge Quantile Regression Models

4.3. Prediction of Extreme Rainfall in 2013

The prediction of extreme rainfall use model M4 which is more consistent in term of RMSE, RMSEP, and correlation of predicted and actual data quantile. The predicted values are also compare to validation data in 2013. Figure 3 shows that the actual rainfall in January is close to the predicted rainfall with lasso regularization in Q(0.75) but is exactly the same the predicted rainfall with ridge regularization. While extreme rainfall in December is between predicted rainfall with lasso regularization in Q(0.75) and Q(0.9) on but is exactly the same the predicted rainfall with ridge regularization

Table 3: Comparison of $RMSEP^{(Percentile)}$ based on Criteria of $\hat{\lambda}$ in Lasso Quantile Regression

Quantile	Year Prediction	RMSEP in				
		Modified Percentile Method	Min Percentile Method	Max Percentile Method	Min Λ	Max Λ
Q(0.75)	2010	$16.85^{(83)}$	$16.28^{(99)}$	$17.27^{(75)}$	$16.28^{(99)}$	$18.62^{(1)}$
	2011	$14.40^{(96)}$	$14.32^{(97)}$	$17.06^{(91)}$	$14.32^{(97)}$	$20.97^{(57)}$
	2012	$15.50^{(92)}$	$15.31^{(82)}$	$42.48^{(100)}$	$14.94^{(68)}$	$42.48^{(100)}$
	2013	$13.88^{(69)}$	$13.57^{(95)}$	$24.29^{(96)}$	$12.49^{(29)}$	$24.29^{(96)}$
Q(0.9)	2010	$23.98^{(93)}$	$22.60^{(79)}$	$55.85^{(95)}$	$17.32^{(26)}$	$55.85^{(95)}$
	2011	$22.20^{(79)}$	$22.20^{(79)}$	$43.26^{(90)}$	$16.63^{(22)}$	$43.26^{(90)}$
	2012	$21.38^{(99)}$	$20.47^{(75)}$	$47.02^{(92)}$	$18.20^{(64)}$	$47.02^{(92)}$
	2013	$17.21^{(80)}$	$15.78^{(82)}$	$49.63^{(100)}$	$10.76^{(17)}$	$49.63^{(100)}$
Q(0.95)	2010	$31.93^{(51)}$	$25.30^{(90)}$	$35.56^{(100)}$	$25.30^{(90)}$	$35.56^{(100)}$
	2011	$46.23^{(8)}$	$21.89^{(86)}$	$58.38^{(81)}$	$21.89^{(86)}$	$58.38^{(81)}$
	2012	$35.26^{(55)}$	$17.66^{(77)}$	$25.76^{(100)}$	$17.66^{(77)}$	$25.76^{(100)}$
	2013	$35.53^{(4)}$	$16.74^{(81)}$	$66.77^{(80)}$	$16.73^{(72)}$	$66.77^{(80)}$

Table 4: Comparison of $RMSEP^{(Percentile)}$ based on Criteria of $\hat{\lambda}$ in Ridge Quantile Regression

Quantile	Year Prediction	RMSEP in				
		Modified Percentile Method	Min Percentile Method	Max Percentile Method	Min Λ	Max Λ
Q(0.75)	2010	$15.20^{(7)}$	$16.27^{(84)}$	$16.83^{(100)}$	$14.61^{(19)}$	$21.58^{(1)}$
	2011	$17.69^{(16)}$	$17.72^{(76)}$	$17.88^{(100)}$	$17.28^{(1)}$	$17.88^{(100)}$
	2012	$17.28^{(47)}$	$17.29^{(78)}$	$17.42^{(99)}$	$17.28^{(22)}$	$18.07^{(1)}$
	2013	$15.94^{(28)}$	$16.24^{(95)}$	$16.63^{(96)}$	$15.45^{(29)}$	$19.31^{(96)}$
Q(0.9)	2010	$17.84^{(1)}$	$18.95^{(75)}$	$21.10^{(100)}$	$17.84^{(1)}$	$21.10^{(100)}$
	2011	$19.28^{(4)}$	$19.68^{(75)}$	$20.55^{(100)}$	$19.21^{(8)}$	$20.55^{(100)}$
	2012	$20.48^{(7)}$	$20.55^{(75)}$	$21.03^{(100)}$	$20.43^{(31)}$	$21.03^{(100)}$
	2013	$15.85^{(10)}$	$17.24^{(82)}$	$18.58^{(100)}$	$15.46^{(17)}$	$18.58^{(100)}$
Q(0.95)	2010	$27.23^{(6)}$	$20.58^{(100)}$	$22.26^{(75)}$	$20.58^{(100)}$	$29.51^{(1)}$
	2011	$28.54^{(19)}$	$23.53^{(100)}$	$24.23^{(76)}$	$23.53^{(100)}$	$32.42^{(1)}$
	2012	$29.74^{(4)}$	$19.41^{(100)}$	$20.66^{(76)}$	$19.41^{(100)}$	$32.94^{(3)}$
	2013	$19.53^{(31)}$	$18.00^{(72)}$	$18.75^{(77)}$	$18.00^{(72)}$	$27.80^{(80)}$

in Q(0.9. Therefore, extreme rainfall in 2013 is better predicted by quantile regression model with ridge regularization than that with lasso regularization.

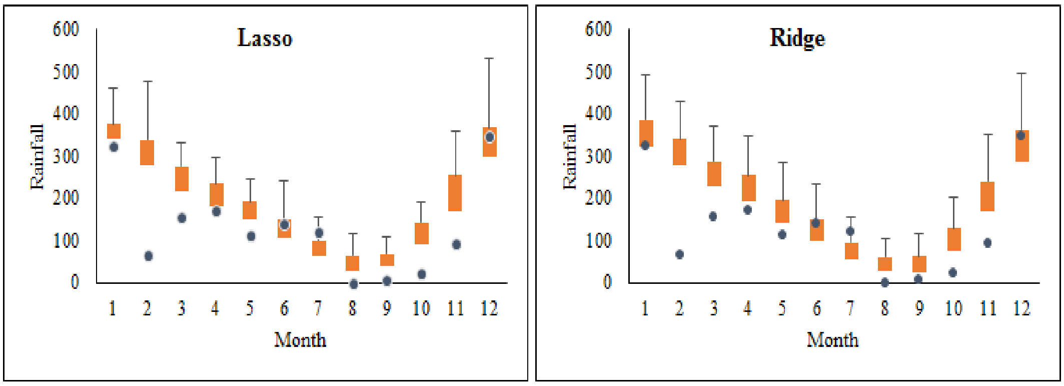

Figure 3: Plot Actual Rainfall 2013 in Lasso and Ridge Quantile Regression Models

Quantile regression with ridge regularization is the best in the prediction of rainfall in 2013 is supported also by the confidence interval (CI) of predicted rainfall. Confidence interval gets from Ì$, \hat{y}_{\tau} \pm (1.96 \times S_{\tau})$ with Ì$, \hat{y}_{\tau}$ is rainfall prediction from M4 in Q(0.75), Q(0.9), and Q(0.95), S_{τ} for standard deviation of all rainfall prediction from M4 Q(0.75), Q(0.9), and Q(0.95). Figure 4 depict a comparison of CI with lasso and ridge regularization. Through the CI plot lasso regularization seen that the higher quantile, the wider the CI of predicted extreme rainfall. This occurs because the value of the standard deviation of the predicted rainfall increases as the increasing the quantile. This case is similar to ridge regularization, but it is not really significant because the standard deviation is too small. The actual quantile data are between upper and lower limit at CI in lasso regularization but close to upper and lower limit at CI in ridge regularization. Therefore, quantile regression with ridge regularization performs better than that lasso regularization because it has the narrow intervals than lasso regularization.

5. Conclusion

The criteria of selection optimum lasso and ridge coefficients based on modified percentile method give good prediction extrem rainfall on either lasso quantile regression and ridge quantile regression. This is indicated by RMSEP is around average of RMSEP for whole coefficients. Value of RMSEP that small and correlation more than 0.9, indicate that the quantile regression with lasso and ridge regularization good at predicting extreme rainfall. In the other hand, the quantile regression with ridge regularization gives better prediction than the lasso regularization. Quantile regression with rigde regularization can capture extreme rainfall in 2013, such as extreme rainfall in January and December at Q(0.75) and Q(0.9).

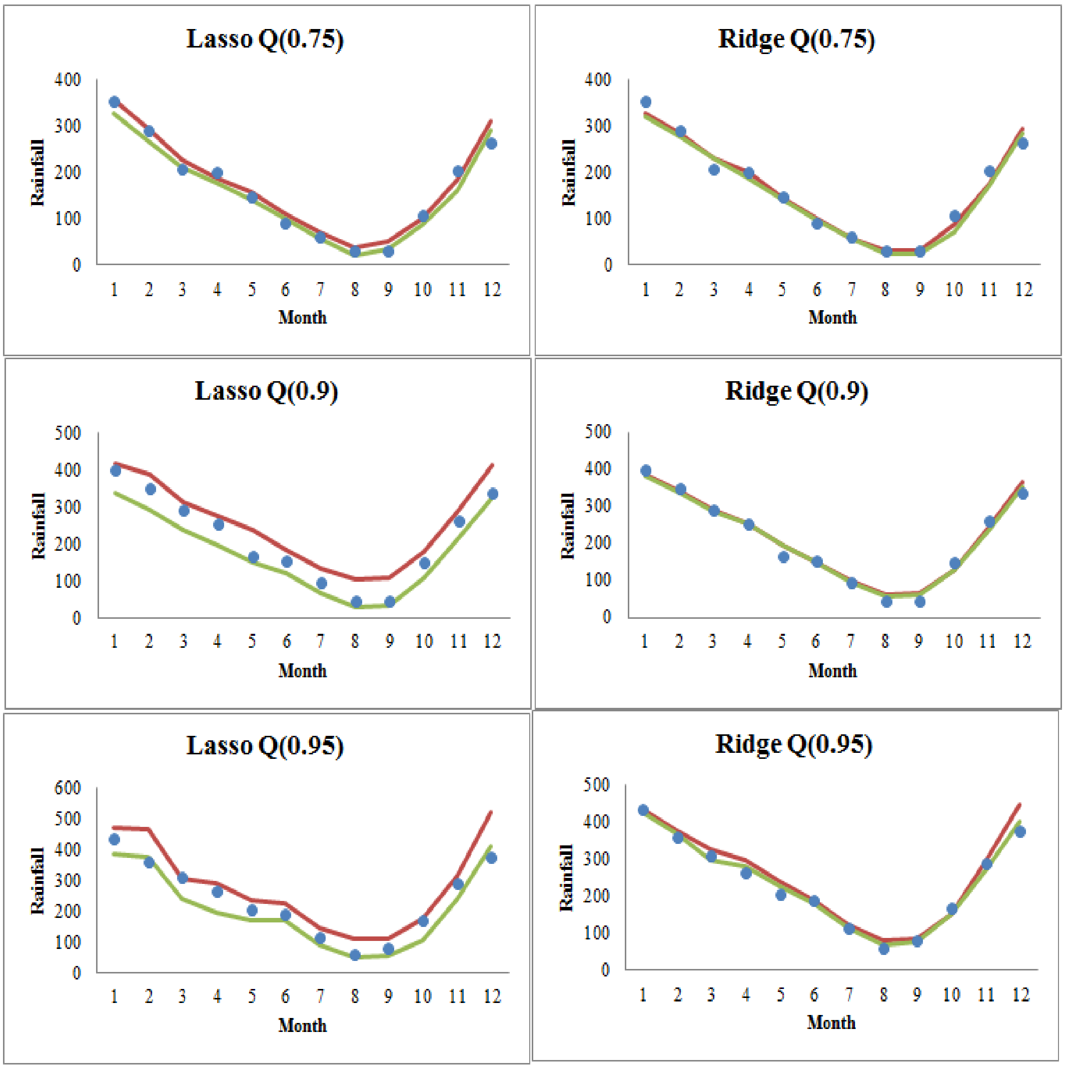

Figure 4: Confidence Interval of Rainfall Prediction in Lasso and Ridge Quantile Regression M4

References

[1] Wigena A.H. and A. Djuraidah. Quantile regression in statistical downscaling to estimate extreme monthly rainfall. *Science Publishing Group*, 2:66–70, 2014.

[2] N.S. Ahmed and E.A.R. Ismail. A comparison between linier regression and lasso quantile regression methods in variable selection. *Current Business and Economics*, 4:69–74, 2015.

[3] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Survey*, 4:40–79, 2010.

[4] P. Buhlman and S.V.D. Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin, 2011.

[5] Tibshirani R. Hastie, T. and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2009.

[6] A.E. Hoerl and R.W. Kennard. Ridge regression: Application to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[7] R. Koenker. *Quantile Regression*. Cambridge, Uk, 2005.

[8] K.V. Lund. The instability of cross-validated lasso. Master's thesis, University of Oslo, Norwegia, 2013.

[9] Koenker R. and K.F. Hallock. Quantile regression. *Economic Perspective*, 15(4):143–156, 2001.

[10] S. Roberts and G. Nowak. Stabilizing the lasso against cross-validation valiability. *Computational Statistics and Data Analysis*, 70:198–211, 2013.

[11] R. Tareghian and P.F. Rasmussen. Statistical downscaling of precipitation using quantile regression. *Hydrology*, 487:122–135, 2013.

[12] R. Tibshirani. Regression and selection via the lasso. *Royal Statistical Society*, 58:267–288, 1996.

[13] Park T. Wang, X. and K.C. Carriere. Variable selection via combined penalization for high-dimensional data. *Computational Statistics and Data Analysis*, 54:2230–2243, 2010.

[14] A.H. Wigena. Partial least square regression for statistical downscaling. *Club BMKG*, 6:10–13, 2011.