Liu-type regression in statistical downscaling models for forecasting monthly rainfall salt as producer regions in Pangkep regency

St Sahriman¹, Anisa¹ and V Koerniawan¹

¹ Statistics Department, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Jl. Perintis Kemerdekaan 10 Tamalanrea Makassar, Sulawesi Selatan, 90245, Indonesia

E-mail: sitisisahriman@yahoo.com

Abstract. Liu-Type Regression (LTR) is one of the statistical methods to overcome multicollinearity in multiple regression models. LTR is the development of Ridge regression and Liu estimator. When there is a strong collinearity, selected k parameter in the ridge regression does not fully overcome the multicollinearity. This study aimed to estimate the rainfall data in Pangkep Regency (as response variable) with LTR approach on Statistical Downscaling (SD) models. Precipitation (as predictor variables) is the result of a simulation of a grid on the Global Circulation Model (GCM). This study uses a size 8x8 grid of GCM (64 predictor variables) over an area of Pangkep Regency so that there is a high multicollinearity. Three dummy variables were determined from k-means cluster technique used as predictor variables to overcome the heterogeneity of residual variance. LTR model with dummy variables are able to explain the diversity of rainfall data properly. The value of $R^2$ produced ranges 85.23% - 88.99% with Root Mean Square Error (RMSE) ranges 117.732-136.377. Validation of the model generates a high correlation value between the actual rainfall and alleged rainfall period of 2017 (about 0.977-0.979). The value of Root Mean Square Error Prediction (RMSEP) produced lower (about 57.625-61.120). SD analysis was also performed with and without the dummy variable in the Ridge regression and LTR. In general, LRT models with dummy (k = 0.652, d = -0.799) is the best model based on the value of $R^2$, RMSE, correlation, and RMSEP.

1. Introduction

Indonesia is one of the owners of the longest coastline in the world. However, it does not make Indonesia into a salt producer that esteemed by the world. South Sulawesi has potential as the largest salt producer, moreover there are some regions that has become the center. Director of Marine and Fisheries Services Marine and Fisheries Ministry Republic of Indonesia, Rianto Basuki, said that Pangkep is the one of ten salt production centers in Indonesia. It caused by good quality of salt. But the fact, their production cannot be fulfilled until now. Production of salt in Pangkep decreased dramatically in 2016, which is only 13,000 tons from the previous year to reach 115,000 tons. This is due to high rainfall. Climate change is quite extreme that makes high-intensity rainfall across Indonesia.

Rainfall is the amount of rainwater that fall on a certain area in a certain time. This climate element has a higher diversity than the temperature and are often used in research as a result of climate change. Climate models need to be developed with a high-resolution global climate circulation regard as superficial precipitation Global Circulation Model (GCM). GCM is a numerical representation of the behavioral climate system and the interaction between components, that are atmosphere, oceans, cryosphere, biosphere, and chemosphere. GCM simulate global climate variables at each grid (measuring ±2.5° or ±300 km²) for each layer of the atmosphere which is used to predict climate patterns in years [1].

Statistical downscaling (SD) is a statistical model of connecting the output of GCM climate variables global-scale with local-scale climate variables [2]. SD has the advantage with easier computation and can be easily applied to a various of outcomes GCM simulations [3]. Precipitation is one of the GCM output variables that can be used to predict the rainfall data of the Earth surface that
has local scale. GCM outputs precipitation generally have large dimensions and has a high correlation (multicollinearity) between grid. Multicollinearity is a bad condition (ill condition) which causes the matrix multiplication of predictor variables (symbolized by \(X'X\)) did not meet the assumptions of classical regression. Multicollinearity causes the error standard of estimating regression parameters to be large and the value of the confidence interval to be wide so that the use of the least squares method (MKT) becomes invalid [4].

In 1970, Hoerl and Kennard introduced a method to overcome the multicollinearity that is method Ridge. This method adds a bias constant positive \(0 < \kappa < 1\) on a matrix of predictor variables. Ridge regression method begins by transforming the response and predictor variables through centralization and scaling procedures [5]. Subsequently, [6] introduced a method that taken from the name itself is the method of Liu estimator. Liu regression method also has a constant value estimate \(d\) as the ridge regression in addressing multicollinearity. However, the constant regression method Liu estimated more easily determined from those of ridge regression method. In 2013, Liu introduced a method with the name of Liu-Type estimator or can be called Liu-Type estimation method using two constants are Ridge constant \((k)\) and Liu constant \((d)\) so that this method can also be called Liu-Type Estimation Method. This is done because the ridge regression of \(k\) parameter is not completely overcome the bad conditions in \(X'X\) [7].

Science studies on the model of rainfall in Indonesia have been carried out by some previous researchers. Among them are [8] using the principle component regression (PCR) and partial least squares regression (PLSR) to predict rainfall on the condition el-nino, la-nina, and normal in Indramayu. Furthermore, [9] utilizes time-shift based on the cross-correlation between the data output GCM precipitation with rainfall data Indramayu in SD using PCR and PLSR. In addition, previous studies researchers improve with the addition of dummy variables based engineering clusters. In the study, the primary method used is the Ridge regression and PCR.

This study discusses the use of SD models with regression method Liu-Type in predicting rainfall data in Pangkep Regency. Prediction of rainfall is also done with the ridge regression method. Dummy variables based k-means cluster technique added as predictor variables in the SD models.

2. Methods

The data used in this research is a global scale of monthly precipitation period 1996-2017 (GCM outputs climate models inter-comparison project that measures 8×8 grid) and rainfall Pangkep Regency local scale (period 1996-2017). The geographical position at 119.57°E until 129.37°W and -14.83°S until 5.17°N in the top around the area Pangkep Regency used as GCM domain. There are 64 predictor variables were used in this study were derived from the domain size 8×8 grid GCM data and three dummy variables. Rainfall Pangkep Regency used as predictor variables in this study. The data used for modeling the period 1966-2016 and data in 2017 for model validation. Multicollinearity identification on the data of precipitation using the variance inflation factors (VIF),

\[
\text{VIF}_j = \frac{1}{1-R_j^2}, \quad j = 1, 2, \ldots, 64
\]

where \(R_j^2\) is determination coefficient of the regression result of predictor variable \(j\) with other predictor variables. In addition, the dummy variables determined from k-means cluster technique.

Data analysis was performed by Liu-Type regression and Ridge regression. The second method begins with selecting a value of bias constants \(k\) and \(d\) to overcome multicollinearity. The initial step in LTR and Ridge regression is to choose the value of \(k\) in order to obtain VIF which is the main diagonal of the matrix below is less than one [10],

\[
(X'X + kI)^{-1}(X'X + kI)^{-2}
\]

where \(X\) is the matrix of predictor variables measures \(n \times (p + 1)\). \(0 < k < 1\) is a constant bias, and \(I\) is the identity matrix size \((p + 1)\). Then the value of \(k\) obtained are used to estimate the value of the constant \(d\) by the following formula [11],

\[
\hat{d} = \frac{\sum_{i=1}^{p}((\lambda_i(\tilde{X}'\tilde{X} - \hat{\beta}\lambda_i^2)))/(\lambda_i^2)}{\sum_{i=1}^{p}((\lambda_i(\tilde{X}'\tilde{X} + \sigma^2)))/(\lambda_i^2)}
\]

Where, \(-\infty < d < \infty\), \(\lambda\) are the eigenvalues of matrix \(X'X\), \(\tilde{X}'\tilde{X}\) is the variance of estimators Ridge regression, \(\hat{\beta}\) is the value of the main diagonal of the matrix \(\tilde{X}'\tilde{X}\), \(\tilde{X} = XQ\). \(Q\) is an orthogonal matrix whose elements are the vector of eigenvalues, and is a response variable matrix \(X'Y\) size \(n \times 1\). Furthermore, the model simulated to some value of bias constant \(k\) and \(d\) that overcome the multicollinearity.

The estimation method Liu-Type written as follows [11], [12],

\[
\hat{\beta}_{\text{Liu}} = \hat{\beta} + d\hat{\beta}
\]

\[
\hat{\beta}_{\text{Liu}} = (X'X + kI)^{-1}(X'Y - d\hat{\beta})
\]
\[ \hat{\mu}_{\text{Liu-Type}} = (X'X + hI)^{-\frac{1}{2}}X'Y - d(X'X + kI)^{-\frac{1}{2}}X'Y \]

Selection of the best models using RMSEP (Root Mean Square Error of Prediction) value and the correlation between the actual value and the estimated value in 2017. The performance of the model is well demonstrated with RMSEP value smaller and higher correlation value. RMSEP value is,

\[ \text{RMSEP} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \]

3. Results and Discussion

(a) Liu-Type Regression Estimation

Parameter estimation method by using Ridge regression of constants bias \( k \) to overcome the multicollinearity problems in the data. However, this constant value is considered still not overcome the problem of multicollinearity in the data optimally. Ridge regression estimation is done by using the form augmentation equation \( U = k^\frac{1}{2} \beta + \tilde{\varepsilon} \) to obtain estimation models of the parameters. Selection of great \( k \) value can cause considerable distances between 0 and \( k^\frac{1}{2} \beta \) thus can lead to substantial bias. Therefore, Kejian Liu obtain a new estimator that is Liu-Ridge by adding shape of augmentation to be,

\[ \begin{bmatrix} \frac{k}{\sqrt{\varepsilon_{\text{L}}}} \beta \end{bmatrix} = \frac{k}{\sqrt{\varepsilon_{\text{L}}}} \beta + \tilde{\varepsilon} \]

so the regression equation into shape following augmentation,

\[ Y = X\hat{\beta} + \varepsilon \]

or can be written into,

\[ Y_R = X_R \hat{\beta} + \varepsilon_R \]

where \( Y_R = \begin{bmatrix} \frac{k}{\sqrt{\varepsilon_{\text{L}}}} X \end{bmatrix}, X_R = \begin{bmatrix} X \end{bmatrix}, \varepsilon_R = \begin{bmatrix} \varepsilon \end{bmatrix} \) \( Y_R \) is the vector on the response variable size \( (n + p) \times 1 \), \( X_R \) the matrix in predictor \( p \) variable size \( (n + p) \times p \), \( \hat{\beta} \) is the regression parameter vector size \( (p \times 1) \), \( \varepsilon_R \) is a residual vector sized \( (n \times 1) \). By applying the method of ordinary least squares (OLS) estimation techniques, the obtained estimator LTR,

\[ \hat{\beta}_{\text{Liu-Type}} = (X'X + hI)^{-\frac{1}{2}}X'Y - d(X'X + kI)^{-\frac{1}{2}}X'Y \]

(b) Multicollinearity Identification

The existence of multicollinearity in the model can be detected either by using a statistical value of VIF. If VIF over 10, it can be concluded that the data contains significant multicollinearity. The result of the calculation of VIF each predictor variable containing the data showed multicollinearity. VIF value ranging between 5.18–2949.55. This means that there is a strong relationship between the data grid GCM precipitation. Therefore, the method used in SD modeling to overcome the problem of multicollinearity in the data GCM precipitation is Liu-Type regression method.

(c) Determination Dummy Variables with K-Means

K-means method is a clustering method to utilize the concept of a centroid or midpoint. There are a number of \( k \) centroid that can be set at the beginning. Figure 1 (a plot between rainfall data component scores and scores of precipitation data components GCM) shows there are 4 groups of rainfall data based on the color group. Group 1 generally occurs from April to October with the intensity of 0-232 mm/month. Group 2 is the group with the intensity of rainfall 233-607 mm/month and generally occurs in February and March. Group 3 is the group with the intensity of rainfall 608-1019 mm/month and generally occur in December and January. Group 4 is a group of rainfall intensity from 1020 to 1541 mm/month and generally occur in December and January. Furthermore, the clustering results are used as the basis for the formation of dummy variables. Thus there are three dummy variables used in SD modeling.
Estimation method of Liu-Type regression (LTR) is a combination of Ridge regression and regression estimators Liu in addressing the problem of multicollinearity. SD modeling with LTR method begins with selecting the value of \( k \geq 0.085 \) so that \( 0.092 \leq d \leq 0.293 \) value thus obtained. With the value of the constant bias, the model LTR without dummy variables showed relatively similar results as in the Ridge regression models (Table 1). Value \( R^2 \) produced ranges 58.67%—60.27% with RSME value is about 221.840—226.261 (Table 1). Thus, LTR models without dummy variables are not able to explain the diversity of the data. Figure 2 provides information that the pattern of the scattered residual of LTR models for every \( k \) and \( d \) value that indicates the tape-shaped heterogeneous residual variance.

### Table 1. The \( R^2 \) and RMSE Ridge Regression Model and Regression Liu-Type Without Dummy

| Method          | \( k \) | \( d \) | \( R^2 \)  | RMSE   |
|-----------------|--------|--------|-----------|--------|
| Ridge Regression (RR) | 0.085  | -      | 62.19%    | 216.408|
|                 | 0.090  | -      | 62.11%    | 216.653|
|                 | 0.100  | -      | 61.95%    | 217.106|
|                 | 0.200  | -      | 60.90%    | 220.093|
|                 | 0.300  | -      | 60.29%    | 221.786|
| Liu-Type Regression (RLT) | 0.085  | 0.092 | 60.27%    | 221.840|
|                 | 0.090  | 0.096 | 60.20%    | 222.048|
|                 | 0.100  | 0.106 | 60.06%    | 222.430|
|                 | 0.200  | 0.197 | 59.17%    | 224.898|
|                 | 0.300  | 0.293 | 58.67%    | 226.261|
The 3rd International Conference On Science
Journal of Physics: Conference Series 1341 (2019) 092021 doi:10.1088/1742-6596/1341/9/092021

Figure 2. Residual Plot of LTR Models

SD modeling with RR method and LTR without dummy variables have a heterogeneous residual patterns. Therefore, dummy variables are added as predictor variables into the SD model. RR and LTR models with dummy variables that referred as RR-dummy and LTR-dummy and able to explain the diversity of better rainfall based on the value of $R^2$ and result of RMSE. The addition of dummy variables in the model SD is able to improve the performance of the model in explaining the rainfall of about 17.90% in the RR-dummy models and approximately 28.72% to the LTR-dummy models (Table 2). RMSE value generated is also lower, both on the model RR-dummy (about 158.301–173.869) and the RLT-dummy models (about 117.732–136.377). In general, LTR-dummy models on $k$ value 0.652 and d amounted to -0.799 best models based on $R^2$ values higher and lower RMSE. Diagnostic of residual models conducted on a RR-dummy and LTR-dummy model. Figure 3 shows the residual variance relatively more homogeneous on RLT-dummy models compared with RLT models.
RR-dummy models and LTR-dummy models are consistent with the higher and lower RMSE of the model without the dummy variables. In general, the LTR-dummy estimates for rainfall data in Pangkep Regency. Model RR-dummy and RLT-dummy produce the RR-dummy method and LTR-dummy can overcome the problem so the results with R² values are

| Method          | R²    | d     | R²   | RMSE   |
|-----------------|-------|-------|------|--------|
| Ridge Regression| 0.652 | -0.799| 88.99%| 117.732|
| (RR-dummy)      | 0.700 | -0.861| 88.51%| 120.283|
| (LTR-dummy)     | 0.800 | -0.980| 87.45%| 125.684|
| (Liu-Type Regression with dummy) | 0.900 | -1.085| 86.35%| 131.085|
| (Liu-Type Regression with dummy) | 1.000 | -1.176| 85.23%| 136.377|

Figure 3. Residual Plot of LRT-dummy Models

Validation of SD Models Method with Liu-Type Regression

RR-dummy method and LTR-dummy can overcome the problem so the results with R² values are higher and lower RMSE of the model without the dummy variables. In general, the LTR-dummy models preferably than the RR-dummy models. LTR-dummy models are also consistent with the results of model validation as the best model. Table 3 presents the correlation values and RMSEP estimates for rainfall data in Pangkep Regency. Model RR-dummy and RLT-dummy produce the alleged use rainfall data is superior than the model without the dummy variables. Prediction of rainfall data using the model RR-dummy and RLT-dummy (RMSEP range 57.625-73.352 and correlations range 0.967-0.979) is better than the RR and LTR models (RMSEP range 89.186-94.384 and correlations range 0.928-0.938).
The model which does not use dummy variables in Table 3 show that the RR model (RMSEP range 89.186-90.146 and correlations range 0.932-0.938) is better than the model of LTR (RMSEP range 91.341-94.384 and correlations range 0.928-0.930), while models involving variables dummy indicate that the LTR models (RMSEP range 57.625-61.120 and correlations range 0.977-0.979) is better than the model of RR (RMSEP range 67.955-73.352 and correlations range 0.967-0.972). In general, LTR-dummy models capable of estimating rainfall data are more accurate than other models. Furthermore, the model LTR-dummy with $k$ value of 0.652 and $d$ value of -0.799 is a model that provides the best results based on alleged of the smallest RMSEP (57.625) and the higher correlation (0.979).

Table 3. Correlation and RMSEP Values of Ridge Regression and Regression Liu-Type Models

| Method               | With Dummy Variables | Without Dummy Variables |
|----------------------|----------------------|-------------------------|
|                      | $k$     | $d$     | $r$     | RMSEP | $k$     | $d$     | $r$     | RMSEP |
| Ridge Regression     | 0.652   | -0.972  | 67.955  | 0.085 | -0.938  | 89.186  |
|                      | 0.700   | -0.971  | 68.842  | 0.090 | -0.937  | 89.464  |
|                      | 0.800   | -0.970  | 70.530  | 0.100 | -0.937  | 89.237  |
|                      | 0.900   | -0.968  | 72.025  | 0.084 | 93.345  |
|                      | 1.000   | -0.967  | 73.352  | 0.300 | 90.146  |
| Liu-Type Regression  | 0.652   | -0.799  | 57.625  | 0.085 | 0.902   | 93.341  |
|                      | 0.700   | -0.861  | 57.883  | 0.100 | 0.906   | 91.503  |
|                      | 0.800   | -0.980  | 58.709  | 0.100 | 0.106   | 91.808  |
|                      | 0.900   | -1.085  | 59.824  | 0.200 | 0.197   | 93.581  |
|                      | 1.000   | -1.176  | 61.120  | 0.300 | 0.293   | 94.384  |

Figure 4 shows that the model of RR and RR-dummy are generally able to capture rainfall patterns in the period of January, February, March, and May. However, the estimated value is lower than the actual value in September and December (Figure 4). In addition, the RR-dummy models capable of capturing precipitation patterns better than RR models. RR-dummy models correctly guessed rainfall data from January to April, June and September. In addition, the distance between actual and predicted values RR-dummy models closer than the RR models.

![Figure 4](image)  
(a) Actual and Predicted Rainfall Plot of (a) RR and (b) RR-dummy Models

Figure 5 is a plot of the actual rainfall and predicted rainfall LTR and LTR-dummy models. LTR models is less able to perform estimation of rainfall followed the pattern of actual rainfall (Figure 5). In addition, the distance between actual and predicted values generated relatively more distant than the LTR-dummy model results. Figure 5 shows that the LTR-dummy models closer to and following the actual rainfall patterns. In general, the model accurately LTR-dummy suspect rainfall data for the period January to September compared with the model of RR, RR-dummy, or LTR. This indicates that the addition of dummy variables in the model can improve the predicted LTR rainfall data.
4. Conclusion

Statistical downscaling modeling with Liu-Type regression method begins with determining the bias constants \( k \) and \( d \) to overcome multicollinearity. The results of the Liu-Type regression modeling with additional dummy variables could explain the diversity of rainfall data properly based on the value of high \( R^2 \) and lower RMSE. Validation of Liu-Type regression model with dummy variables are preferable than Ridge regression and Liu-Type regression models without dummy variables. The range of correlation value is 0.977-0.979 and RMSEP range is 57.625-61.120. In general, Liu-Type regression model with dummy variables on the \( k \) value of 0.652 and \( d \) value of -0.799 is the best model.

Acknowledgement

We thank PDPU of LP2M (Institute for Research and Community Service) Hasanuddin University as the internal research grant provider of Hasanuddin University in 2019. We also thank all staff from the Statistics Department and the Faculty of Mathematics and Natural Sciences who have provided all the facilities for the completion of this research.

References

[1] Wigena AH. 2006. Statistical modeling downscaling with regression of projection persuit for monthly rainfall forecasting [dissertation]. Bogor (ID): Bogor Agricultural Institute.
[2] Fernandez E. 2005. On the influence of predictors area in statistical downscaling of daily parameters. *Norwegia Meteorological Institute*. 9:1-21.
[3] Sutikno. 2008. Statistical downscaling of GCM outputs and their use for predicting rice production [dissertation]. Bogor (ID): Bogor Agricultural Institute.
[4] Montgomery DC, Peck EA. 1992. *Introduction To Linear Regression Analysis Second Edition*. New York: John Willey and Sons Inc.
[5] Jerry L. Kaysville H, Utah. 2007. *Ridge Regression*. U.S: NCSS.
[6] Liu K. 1993. A New Class of Biased Estimate in Linier Regression. *Communications in Statistics*. 22(1): 393-402.
[7] Li Y, Yang H. 2012. A New Liu-Type Estimator in Linear Regression Model. *Stat Papers*. 53:427-437.
[8] Estinintyas W, Wigena AH. 2011. Statistical statistical downscaling with principle component regression and partial least squares regression for rainfall prediction in el nino, la nina, and normal conditions. *Journal of Meteorology and Geophysics*. 12(1):65-72.
[9] Sahriman S. 2014. Statistical Downscaling Model with Time Lag of Global Circulation Model to Forecast Rainfall [Thesis]. Bogor (ID): Bogor Agricultural Institute.
[10] Layla RDNK, Eka FP. 2016. *Different and Ridge Regression*. Surabaya: Institute of Technology Sepuluh Nopember.

[11] Liu K. 2013. Using Liu-type estimator to combat multicolinearity. *Biostatistics and Bioinformatics Core*, USA.

[12] Kurnaz FS, Akay KU. 2015. A New Liu-type Estimator. *Stat Papers*. 56:495-517