Ridge Robust Regression Analysis on Measurement of Non-Invasive Blood Glucose Levels

Linda Rassiyanti[1], Erfiani[2], Indahwati[2]

[1] Student of Department Statistics, IPB University, Bogor, 16680, Indonesia
[2] Lecturer of Department Statistics, IPB University, 16680, Indonesia

DOI: 10.29322/IJSRP.10.06.2020.p102109
http://dx.doi.org/10.29322/IJSRP.10.06.2020.p102109

Abstract- Calibration method is one of the way that can be used to analyze the relationship between invasive and non-invasive blood glucose levels. The problems which often occur in the calibration method of blood glucose levels are multicollinearity and outliers. One method that can be used to overcome this problem is the ridge robust-MM. This method is a combination of robust regression with MM estimators that are robust to outliers and ridge regression to overcome multicollinearity. This study aimed to compare modulation 0 to 90 and modulation 50 to 90 using the ridge robust-MM regression. The data used in this study were invasive and non-invasive blood glucose collected in 2019 and consisted of 74 respondents. The results show that modulation 0 to 90 had bigger coefficient determination (R^2) value and lower Root Mean Square Error (RMSE) value than modulation 50 to 90, which was 47.62% and 5.71e-02 respectively. The RMSEP value of modulation 0 to 90 was 1.66e-01.

Index Terms- diabetes mellitus, non-invasive, calibration, ridge robust-MM regression.

I. INTRODUCTION

Poor lifestyle, can cause hyperglycemia. Hyperglycemia is a high level of blood sugar in the body. This condition can cause someone stricken with Diabetes Mellitus (DM). Healthy lifestyle and early detection is needed as an effort to improve the DM. Detection of blood sugar can be done by invasive and non-invasive method. The invasive method is done by injuring the limbs, this way can cause discomfort for patients [1]. The non-invasive method can detection of blood sugar levels without injuring the patient. The output of non-invasive method is spectrum of residual intensity on the time domain while the output of invasive method is the result of laboratory tests conducted by Prodia. Residual intensity have 10 modulation, namely modulation 0 to 90. Modulation is the lighting level setting on non-invasif method.

Calibration method is one of the way that can be used to analyze the relationship between invasive and non-invasive blood glucose levels. The problem that often occurs in calibration modeling are multicollinearity and outliers. One method that can be used to overcome this problem is the ridge robust-MM. This method is a combination of robust regression with MM estimators that are robust to outliers and ridge regression that can overcome multicollinearity. This study aimed to compare modulation 0 to 90 and modulation 50 to 90 using the ridge robust-MM regression.

II. LITERATURE REVIEW

A. *Outliers*

Outliers are values that are far different from the other values. The existence of outliers can cause various problems, one of which is not fulfilling the normality assumption. [2] The existence of outliers will also change the conclusions made by researchers because the parameter estimator values are biased. There are several methods and values used to detect outliers. The boxplot is one of the method to see if there are outliers in the data. In addition, there are also outliers that can be seen based on the value of DfFITS, Leverage Values, Cook's Distance, and DfBETA [3].

B. *Multicollinearity*

Multicollinearity means there is a linear relationship between some or all of the independent variables in the model [4]. Multicollinearity can only be found in multiple linear regression. Multicollinearity can be detected using the Variance Inflation Factor (VIF) value [5].

$$VIF_{(j)} = \frac{1}{(1-R_j^2)} \ , j = 1,2, \dots , k$$

when R_j^2 is the coefficient of determination obtained from the independent variable (X_j) which is regressed with other independent variables.

If there is multicollinearity, it will have the following effects:

This publication is licensed under Creative Commons Attribution CC BY.

1. The partial coefficient of regression is not measured precisely, thus making the standard error value large.
2. Estimated ordinary least square parameters and standard deviations will be very sensitive to changes.
3. The confidence interval value of the regression coefficient will widen, so that it tends to accept H_0 or there is no statistically significant regression coefficient [6].

C. Robust Regression

The robust regression is a regression analysis method that is used when distribution of error is not normal or there is an outlier. The robust regression can overcome data with outliers without eliminating the outliers [7]. Estimation methods in robust regression include M (maximum likelyhood type) estimation, LTS (least trimmed squares) estimation, S (scale) estimation, and MM (method of moment) estimation. This study used MM estimator which is a method that combines S estimation (estimation with high breakdown point) and M estimation introduced by Yohai.

D. Ridge Regression

The ridge regression gives a biased estimate of regression coefficient by modifying the least squares method to get minimum variance by adding a constant k for stabilizing the coefficient [8]. If value of k is 0 then $\beta_{ridge} = \beta_{OLS}$ and if value of k> 0, β_{ridge} is a biased but more stable estimator. The modification is done by adding the bias constant k to the diagonal matrix $\mathbf{X^TX}$ [9]. The independent variable and the dependent variable in the ridge regression are transformed into a standard form (standardization). The equation of ridge regression is:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}} = (\boldsymbol{X^TX} + k\boldsymbol{I})^{-1}\boldsymbol{X^TY}, \quad k > 0$$

E. Ridge Robust Regression

The ridge robust MM is a combination of robust regression with MM estimators that are robust to outliers and ridge regression that can overcome multicollinearity [10]. The result of ridge robust regression will be stable and resistant to outliers. The formula for estimating the ridge robust regression parameter is as follows:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{RR}} = (\boldsymbol{X^TX} + k\boldsymbol{I})^{-1}\boldsymbol{X^TX}\widehat{\boldsymbol{\beta}}_{\boldsymbol{RobustMM}}$$

One of the ways to get the k value is use the method introduced by Hoerl, Kennard and Balwin in 1975. The formula for finding the k value is:

$$k = \frac{p\hat{\sigma}^2_{RobustMM}}{\hat{\beta}'_{\text{RobustMM}}\hat{\beta}_{\text{RobustMM}}}$$

III. DATA AND METHODOLOGY

A. Data

The data used in this study were invasive and non-invasive blood glucose collected in 2019. The data was collected at Tanah Sereal District, Bogor City and consisted of 74 respondents. The independent variable was invasive blood glucose measurement from laboratory tests conducted by Prodia and the dependent variable was non-invasive blood glucose measurement.

The tool of non-invasive measurement was designed with 5 replications. Each replications contains 10 modulation, namely modulation 0 to 90. Modulation is the lighting level setting on non-invasif method. So, for modulation 0 to 90 had 50 independent variables, because there were 10 modulations with 5 replications. Modulation 50 to 90 had 25 independent variables, because there were 5 modulations with 5 replications.

B. Methods

The analysis procedure in this study were:

1. Testing outliers and multicollinearity.
2. Calculate the value of β_{MM} using the robust-MM regression.
3. Calculate the value of k.
4. Calculate the value of $\beta_{RR.}$
5. Measuring the goodness of the model using the values of R^2, RMSE, and RMSEP.
6. Draw a conclusion.

IV. RESULT AND DISCUSSION

Data with modulation 0 to 90 had 50 variables while modulation 50 to 90 had 25 variables. Each modulation was centralized and scaled, and divided into 2 parts namely training data (80%) and testing data (20%). The sampling was repeated 100 times.

A. Outliers and Multicollinearity

The highest blood glucose level was 614 mg/dL and the lowest glucose level was 69 mg/dL. [11] Blood glucose levels are divided into three, normal (<100 mg/dL), pre-diabetes (≥100 mg/dL to <126 mg/dL) and diabetes (≥ 126 mg/dL). Thirty five respondents had blood glucose levels below 100 mg/dL which were categorized as normal. While 11 respondents had blood glucose levels between

This publication is licensed under Creative Commons Attribution CC BY.

100 mg/dL to 126 mg/dL which were categorized as prediabetes and the 28 respondents had blood glucose levels above 126 mg/dL that were categorized as diabetes.

Outliers in this study was respondents 15, 28, 29, 33, 40, 48, 64, and 76 which had blood glucose levels 282 mg / dL, 274 mg / dL, 328 mg / dL, 258 mg / dL, 303 mg / dL, 319 mg / dL, 256 mg / dL, and 614 mg / dL respectively. Outliers in the data will make mistakes, variance in the data, and intervals have a wider range [12]. In this study, outliers are still included in the analysis process using robust methods to outliers.

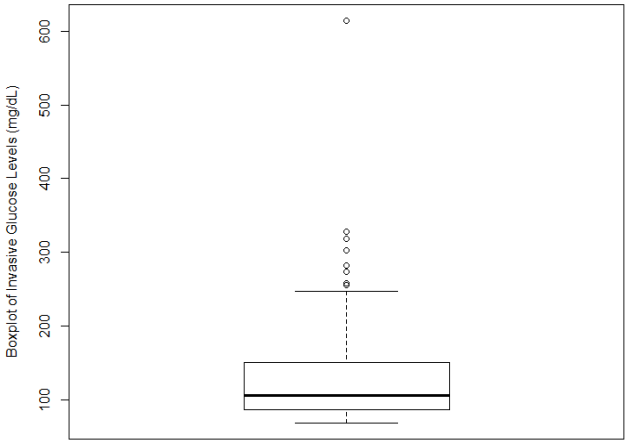

Figure 1: Boxplot of Invasif Blood Glucose Levels

Detection of multicollinearity can use the VIF value of each independent variable. The VIF value of modulation 0 to 90 presented in table 1 and the VIF value of modulation 50 to 90 presented in table 2. If the value of VIF is more than 1 then it is said that there is low multicollinearity and if the value of VIF is more than 5 then there is a high multicollinearity [13]. Based of Table 1, all VIF values were higher than 5, which means there was a high multicollinearity problem in modulation 0 to 90 and based of Table 2, almost all the value of VIF for each independent variable were higher than 5 too. So, modulation 50 to 90 had high multicollinearity problem too.

Tabel 1: The VIF value of modulation 0 to 90

Variable	VIF Value	Variable	VIF Value	Variable	VIF Value
X_1	45.69	X_{18}	16.78	X_{35}	30.20
X_2	26.30	X_{19}	30.47	X_{36}	78.03
X_3	17.31	X_{20}	54.38	X_{37}	21.38
X_4	13.31	X_{21}	30.19	X_{38}	29.29
X_5	18.69	X_{22}	64.99	X_{39}	12.54
X_6	28.03	X_{23}	29.26	X_{40}	17.36
X_7	7.96	X_{24}	23.90	X_{41}	43.35
X_8	26.16	X_{25}	31.81	X_{42}	27.20
X_9	20.40	X_{26}	26.86	X_{43}	17.20
X_{10}	27.47	X_{27}	26.11	X_{44}	60.15
X_{11}	32.99	X_{28}	23.42	X_{45}	35.40
X_{12}	59.93	X_{29}	34.09	X_{46}	39.03
X_{13}	22.68	X_{30}	36.21	X_{47}	24.25
X_{14}	24.12	X_{31}	31.71	X_{48}	40.35
X_{15}	29.40	X_{32}	39.11	X_{49}	40.52
X_{16}	27.89	X_{33}	17.39	X_{50}	17.38
X_{17}	29.60	X_{34}	25.44		

Tabel 2: The VIF value of modulation 50 to 90

This publication is licensed under Creative Commons Attribution CC BY.

Variable	VIF Value	Variable	VIF Value
X_1	12.74	X_{14}	10.00
X_2	3.33	X_{15}	12.46
X_3	10.13	X_{16}	20.35
X_4	7.86	X_{17}	12.32
X_5	10.91	X_{18}	11.28
X_6	12.69	X_{19}	8.80
X_7	15.86	X_{20}	9.65
X_8	11.48	X_{21}	15.31
X_9	16.41	X_{22}	12.63
X_{10}	16.97	X_{23}	14.84
X_{11}	14.66	X_{24}	14.52
X_{12}	12.29	X_{25}	9.48
X_{13}	10.30		

B. Ridge Robust-MM

After getting a parameter estimation of robust-MM regression, then this value is used to calculate the value of k. The value of k for modulation 0 to 90 was 0.01. This k value is useful for stabilizing the estimator coefficient in ridge robust. The parameter estimation of robust-MM regression and k value was used to estimate parameter of the ridge robust-MM. The Goodness of model was presented in Table 3. The R^2 values was 4.76e-01 or equal to 47.62%. This means that the variance of the dependent variable can be explained 47.62% by model, while the remaining 53.38% cannot be explained by model. The RMSE value was 5.71e-02 while the RMSEP value was 1.66e-01.

Table 3: Goodness of model for modulation 0 to 90

Statistics	Training Data		Testing Data
	R^2	RMSE	RMSEP
Mean	4.76e-01	5.71e-02	1.66e-01
Variance	1.97e-02	2.87e-04	1.96e-03

The value of k for modulation 50 to 90 was 0.11. This k value is useful for stabilizing the estimator coefficient. The goodness of model was presented in Table 4. Based Table 4, the R^2 values was 1.99e-01 or equal to 19.88%. This means that the variance of the dependent variable can be explained 19.88% by model, while the remaining 81.22% cannot be explained by model. The RMSE value was 1.02e-01 while the RMSEP value was 1.19e-01.

Table 4: Goodness of model for modulation 50 to 90

Statistics	Training Data		Testing Data
	R^2	RMSE	RMSEP
Mean	1.99e-01	1.02e-01	1.19e-01
Variance	5.10e-02	2.10e-04	1.66e-03

V. CONCLUSIONS AND RECOMMENDATIONS

A. Conclusions

The results of this study is modulation 0 to 90 had bigger coefficient determination (R^2) value and lower Root Mean Square Error (RMSE) value than modulation 50 to 90, which was 47.62% and 5.71e-02 respectively. The RMSEP value of modulation 0 to 90 was 1.66e-01.

B. Recommendations

This publication is licensed under Creative Commons Attribution CC BY.

In future studies of non-invasive blood glucose levels, it is better to use modulation 0 to 90 in the analysis. Calibration modeling using ridge robust regression can be used in future studies using the other parameter estimator.

REFERENCES

[1] Satria E, Wildian, "Rancang bangun alat ukur kadar gula darah non-invasif berbasis mikrokontroler AT89S51 dengan mengukur tingkat kekeruhan spesimen urine menggunakan sensor *Fotodioda",* Journal of Fisika Unand, 2013, pp. 40-47.
[2] Drapper NR and Smith H, *Applied Regression Analysis*, New York (US): John Wiley and Sons, Inc., 1998.
[3] Soemartini, *Pencilan (Outlier)*, Bandung (ID): Universitas Padjajaran, 2007.
[4] Gujarati DN and Porter DC, *Basics Econometrics*, 5th ed, New York (US): McGraw-Hill/Irwin, 2009.
[5] Montgomery DC and Runger GC, *Applied Statistics and Probability For Engineers*, 5th ed, New York (US): John Wiley and Sons, Inc., 2011.
[6] Jolliffe IT, *Principal Component Analysis*, 2nd ed, New York (US): Springer-Verlag, 2002.
[7] Chen C, "Robust regression and outlier detection with the ROBUSTREG procedure. Statistics and Data Analysis", North Carolina: SAS Institute, 2002, pp. 265-27.
[8] Mardikyan S and Cetin E, "Efficient choice of biasing constant for ridge regression", *Int. J. Contemp. Math. Sciences*, 2008, pp. 527-536.
[9] Dereny M and Rashwan NI, "Solving multicollinierity problem using ridge regression models", *Int. J. Contemp. Math. Sciences*, 2011, pp. 585-600.
[10] Samkar H and Alpu O, "Ridge regression based on some robust estimators", *Journal of Modern Applied Statistical Methodes*, 2010, pp. 17.
[11] [ADA] American Diabetes Association (USA), "Diagnosis and classification of diabetes mellitus", *Diabetes Care*, 2014, pp. 81-90.
[12] Ismah, Wigena AH, and Djuraidah A, "Pendekatan regresi kuadrat terkecil partial robust dalam model kalibrasi". *Statistics and Computing Forum*, 2009, pp. 34-41.
[13] Daoud IJ, "Multicollinierity and regression analysis", *Journal of Physics*: Conf. Series 949, 2017.

This publication is licensed under Creative Commons Attribution CC BY.