The Calibration Model for Predicting Non – Invasive Blood Glucose Levels by Using the Principal Component of Quantile Regression

Siti Arita Novia[*], Erfiani[*], Aji Hamim Wigena[*]

[*] Department of Statistics, IPB University

DOI: 10.29322/IJSRP.10.11.2020.p10722
http://dx.doi.org/10.29322/IJSRP.10.11.2020.p10722

Abstract- Diabetes mellitus (DM), a chronic metabolic disorder, is caused by which the pancreas does not produce enough insulin (a hormone that regulates blood glucose) or of which the body cannot use the insulin which is produced effectively. If it is not immediately prevented or addressed, DM will induce to complications leading to death. One of the preventive actions is to check blood glucose levels regularly. Checking blood glucose levels is usually conducted in invasive ways such as by injection and a glucometer which injure the body, take a long time and spend a big expense. Based on those phenomena, the IPB bio marking team is motivated to develop a non-invasive blood glucose monitoring device that is non-injurious for the body. The development of this blood test requires the calibration analysis model. This model can be used to analyze the relationship between invasive and non-invasive blood glucose levels. The principal component of quantile regression calibration model was utilized in this case. The data used in this study were invasive and non-invasive blood glucoses collected from 118 respondents in 2017. The result of the study showed that the model with a quantile of 0,70 was good for predicting non-invasive blood glucose levels with RMSEP value of 0,0852.

Index Terms- calibration model, diabetes mellitus, non-invasive, quantile regression

I. Introduction

Diabetes mellitus (DM), a chronic metabolic disorder, is caused by which the pancreas does not produce enough insulin (a hormone that regulates blood glucose) or of which the body cannot use the insulin which is produced effectively. As a result, there is an increase in the concentration of glucose in the blood (hyperglycemia). If it is not immediately prevented or addressed, DM will induce to complications leading to death. Indonesia is the sixth-biggest country having problem with DM in the world after China, India, the United States, Brazil, and Mexico. In addition, [9] states that the number of adults from the world's population suffering from diabetes is estimated to increase from 150 million in 2000 to 300 million in 2025. This estimate makes diabetes one of the global health problems with the fastest-growing emergency in the 21st century [3]. This problem requires preventive actions, including checking blood glucose levels regularly [9].

Furthermore, checking blood glucose levels is usually conducted in invasive ways such as by injection and a glucometer which injure the body, take a long time and spend a big expense. Based on those phenomena, the IPB bio marking team is motivated to develop a non-invasive blood glucose monitoring device that is non-injurious for the body. The development of this blood test requires the calibration analysis model. According to [5], the multiple variable of calibration model has a function of finding the relationship both measurement units that can be obtained through a relatively easy or inexpensive process and measurement units that require a long time and are expensive. The goal of calibration model is to find the good relationship so that expensive measurements can be predicted quickly, are highly accurate but clearly inexpensive, and spend less money and time. The calibration model in this developed non-invasive device is purposed to determine the relationship between invasive and non-invasive measurements in predicting blood glucose levels. [6] conducted a calibration model by comparing the principle component of regression method, the partial least squares regression and the support vector regression in predicting non-invasive blood glucose levels. The study showed that there were outlier data on the results of measuring invasive blood glucose levels. However, the outlier data were omitted since they have different roles. According [7], outliers can play a positive or negative role in the decision-making process, depending on the type of the case. The outlier data in this case cannot be simply omitted, because there is a suspicion of which a patient has a blood glucose level that is extremely higher than the normal blood glucose levels.

Moreover, in this case the quantile regression method can be applied since the method is flexible in data modeling, especially when extreme values are such important issue [1]. However, calibration modeling requires to reduce the variables first because in general, the calibration data is high multi collinear of explanatory variables [8]. The method that can be used to solve this problem is the principal component analysis. The model of the principal component analysis has the highest sensitivity values so that it can increase the accuracy of the model [4]. This study aims to build a calibration model between invasive and non-invasive blood glucose measurements in predicting non-invasive blood glucose by the principal component of quantile regression.

II. Method

A. Data

The data used in this study are primary data as parts of the research "Calibration Model Through a Continuous Wavelet Transformation Statistical Model, Robust Partial Least Squares Regression, Gaussian Process Regression for developing the Non - Invasive Monitoring System for Patients with High Blood Glucose Levels" by the Non - Invasive bio marking team, Bogor Agricultural University (IPB). Data was taken from April 2016 to January 2017 with a total of 118 respondents who were students from various departments at IPB. Variable Y is the result of invasive measurement in the form of blood glucose levels (mg / dL). Variable X is the result of non-invasive measurement in the form of a spectrum of residual light intensity toward the time domain.

How the non-invasive way works is by irradiating the fingers with an infrared wavelength of 1600 nm (Figure 1). The result of this irradiation is the intensity of light of which it is passed by the limbs. The light is captured by the sensor in the form of a continuous analog voltage value. This value is transformed into a discrete digital voltage value through the Analog-Digital Converter (ADC) which is then processed by the Fast Fourier Transform (FFT) algorithm to produce a time-domain spectrum toward light intensity displayed on a Liquid Crystal Display (LCD). The output of this device is known as the residual light intensity toward the time domain.

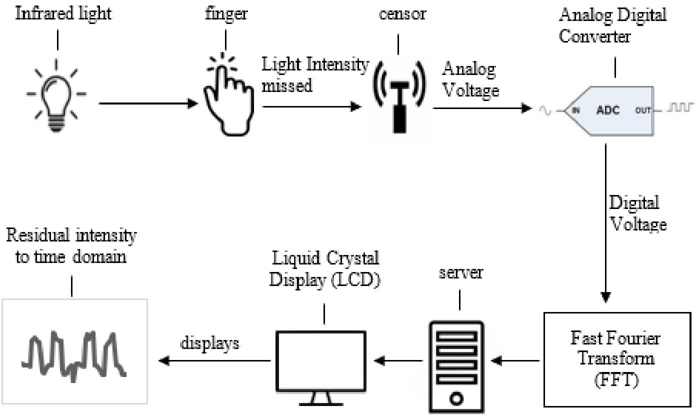

Figure 1: An illustration of how a non-invasive device works

The level of light (modulation) on non-invasive measuring instruments is well designed to modulate 0, 10, 20, 30, 40, 50, 60, 70, 80, and 90 with which each modulation is repeated five times. Modulation 0 means the lights off; modulation 10 is 10.85% of the lights on, the rest is off; modulation 20 is 21.70% of the lights on, the rest of the lights is off; modulation 30 is 32.55% of the lights on, the remaining lights are off; modulation 40 is 43.44% of the lights on, the rest lights are off; modulation 50 is 54.25%of the lights on, the rest is off; modulation 60 is 65.10% of the lights on, the rest lights are off; modulation 70 is 75.95% of the lights on, the rest is off; 80% modulation is 86.80% of the lights on, the rest is off; modulation 90 is 97.65% of the lights on, the rest is off.

B. Procedure of Analysis

Data analysis procedures were determined by using Microsoft Excel and R 4.0.2 software using the *caret*, *rpart*, and *quantreg* packages. The analysis steps in this research are:

1. Exploring the results of invasive blood glucose measurements.
   a. Accomplishing the descriptive statistical analysis.
   b. Identifying any extreme (outlier) blood glucose levels.
2. Preprocessing data which are obtained from non-invasive blood glucose measurements by summarizing the spectrum domain in each period that has been well organized with a time of 500 ms. The immensity formula used is the immensity of the trapezoid with the following equation:

$$X = \sum_{i=0}^{n_t - 1} \frac{(t_{i+1} - t_i)(w_i - w_{i+1})}{2}$$

Information:
X = The immensity of the period curve for each replication (Cdns)
t_i = i-th time domain(ns)
w_i= residual value of the i-th light intensity (Cd)
n_t = number of time domain ranges

The summarizing is done because each replication in the modulation has a different number of residual points of light intensity toward the time domain for each respondent (Figure 2).

http://dx.doi.org/10.29322/IJSRP.10.11.2020.p10722

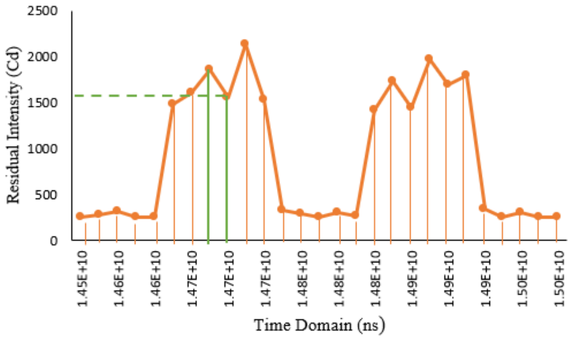

Figure 2: An illustration of a summary of the area of a trapezoid in each period

3. Accomplishing the principal component analysis using the entire original variable data to obtain several principal components which are selected to represent the diversity of data based on the proportion of the cumulative diversity of 95%.
4. Dividing the data both from the measurement of invasive blood glucose levels (y) and from the third procedure into two randomly selected parts, which are 80% for training data and 20% for testing data.
5. Building a quantile regression model at quantiles 0.60, 0.65, 0.70, 0.75, and 0.80 between y and the selected main components using training data.
6. Making predictions according to the model obtained in the fifth step using the testing data.
7. Conducting a test for the virtue model which is formed by using Root Mean Square Error Prediction (RMSEP). The smaller the value obtained, the more appropriate the model is formed [5].
8. Selecting the best principal component quantile regression model based on the seventh step for non-invasive prediction of blood glucose levels.

III. RESULT

A. Data Exploration

The measurement of invasive blood glucose levels was accomplished at the Prodia Laboratory, which is located at Jl. Jend. Sudirman No. 38 B, Bogor 16143. The data from the measurement are presented in (Figure 3), showing that most of the respondents had blood glucose levels less than 90 mg / dL with a small variety and that there were several outliers on the data. The blood glucose levels of respondents who became outliers included 95 mg / dL, 96 mg / dL (three respondents), 103 mg / dL (two respondents), 104 mg / dL, 105 mg / dL, 115 mg / dL, 116 mg / dL, 123 mg / dL and 276 mg / dL. The results of the descriptive statistics show that the average blood glucose level of the respondents was 82.64 mg / dL. In addition, the lowest blood glucose level was 67.00 mg / dL, while the highest blood glucose level was 276.00 mg / dL. This research was conducted without discarding outlier data.

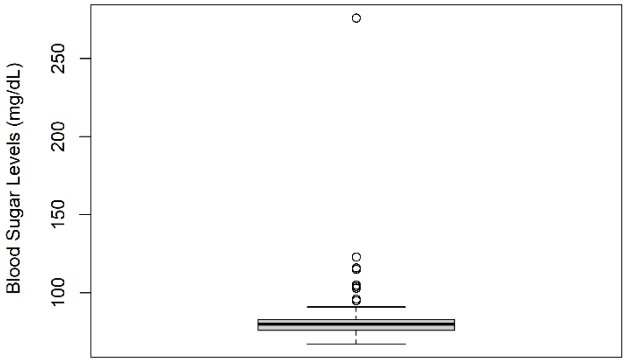

Figure 3: *Box plot* of invasive blood sugar levels

The exploration data of non-invasive measurement (Figure 4) is the 79th respondent's data of the third test. The graph shows that modulation 0 to modulation 40 tends to be more constant than modulation 50 to modulation 90. Kania (2020) states that modulation 50 to modulation 90 has a significant residual value, so this study only uses modulation 50 to modulation 90.

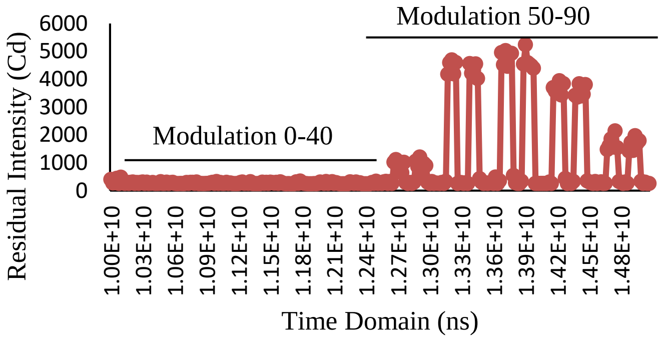

Figure 4 :Graph of non-invasive measurement

If modulation 50 to modulation 90 is re-visualized as presented in (Figure 5), it is clear that each modulation has a number of residual points of different intensity toward the time domain so that the data is summarized using a trapezoidal immensity.

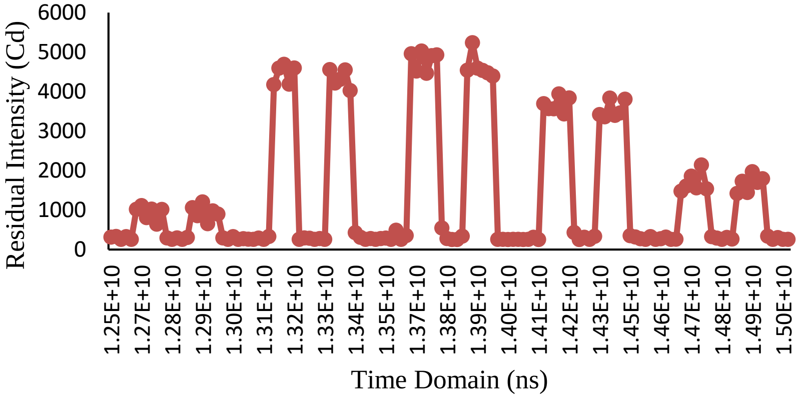

Figure 5 :Visualization of non-invasive 50-90 measurement modulation

There were 25 variables obtained from the result of summarizing the period immensity. This is because each of the 50, 60, 70, 80, and 90 modulations was repeated five times.

B. Quantile Regression Model

The first step before modeling is to reduce the dimensions of the variables with principal component analysis (PCA). A total of 25 original variables obtained from summarizing the immensity of the period were reduced in dimensions. The PCA results show that there were 11 principal components (PC) to represent the diversity of the data obtained from the proportion of the cumulative diversity of 95%. Furthermore, 11 PC and Y variable data were modeled with quantiles 0.60, 0.65, 0.70, 0.75, and 0.80 by using 80% training data. The results of the model built are obtained as follows:

$$\begin{aligned}
y_{\tau\,=\,0,60} &= 81{,}2404 - 0{,}3117\ \mathrm{PC1} + 1{,}4379\ \mathrm{PC2} + 0{,}7977\ \mathrm{PC3} - 0{,}2064\ \mathrm{PC4} - 2{,}1938\ \mathrm{PC5} - 0{,}9506\ \mathrm{PC6} - 0{,}8697\ \mathrm{PC7} \\
&\quad - 2{,}0940\ \mathrm{PC8} + 0{,}2673\ \mathrm{PC9} - 2{,}3708\ \mathrm{PC10} - 2{,}4591\ \mathrm{PC11} \\
y_{\tau\,=\,0,65} &= 82{,}1232 - 0{,}2861\ \mathrm{PC1} + 2{,}0448\ \mathrm{PC2} + 0{,}6990\ \mathrm{PC3} - 0{,}0163\ \mathrm{PC4} - 1{,}6540\ \mathrm{PC5} - 1{,}2220\ \mathrm{PC6} - 0{,}9565\ \mathrm{PC7} \\
&\quad - 2{,}1353\ \mathrm{PC8} - 0{,}2230\ \mathrm{PC9} - 2{,}8943\ \mathrm{PC10} - 1{,}9122\ \mathrm{PC11} \\
y_{\tau\,=\,0,70} &= 83{,}1845 - 0{,}2486\ \mathrm{PC1} + 2{,}6386\ \mathrm{PC2} - 0{,}1908\ \mathrm{PC3} - 1{,}6354\ \mathrm{PC4} - 0{,}7232\ \mathrm{PC5} - 1{,}1123\ \mathrm{PC6} + 0{,}8370\ \mathrm{PC7} \\
&\quad - 2{,}3713\ \mathrm{PC8} + 1{,}4850\ \mathrm{PC9} - 2{,}7538\ \mathrm{PC10} + 0{,}3477\mathrm{PC11} \\
y_{\tau\,=\,0,75} &= 85{,}2691 - 0{,}2006\ \mathrm{PC1} + 3{,}0334\ \mathrm{PC2} - 0{,}3078\mathrm{PC3} + 1{,}7545\ \mathrm{PC4} + 0{,}7908\mathrm{PC5} - 2{,}6889\ \mathrm{PC6} - 0{,}6438\ \mathrm{PC7} \\
&\quad - 2{,}1602\mathrm{PC8} + 0{,}3755\ \mathrm{PC9} - 2{,}5799\ \mathrm{PC10} + 1{,}7203\mathrm{PC11} \\
y_{\tau\,=\,0,80} &= 86{,}2339 - 0{,}1417\ \mathrm{PC1} + 3{,}1772\ \mathrm{PC2} + 0{,}2389\mathrm{PC3} + 0{,}5759\ \mathrm{PC4} + 0{,}7785\mathrm{PC5} - 3{,}8377\ \mathrm{PC6} - 0{,}4624\ \mathrm{PC7} \\
&\quad - 2{,}5204\ \mathrm{PC8} - 0{,}3533\ \mathrm{PC9} - 1{,}4551\ \mathrm{PC10} + 1{,}6622\mathrm{PC11}
\end{aligned}$$

C. Goodness of Fit

The models that had been built were then predicted by using the 20% test data. The prediction results of these models were tested by using the Root Mean Square Error of Prediction (RMSEP) value. The model with the lowest RMSEP value was chosen as the principal component quantile regression calibration model for the non-invasive prediction of blood glucose levels. Table 1 shows the RMSEP values at various quantiles.

Table 1 The results of the model virtue test

Quantile	RMSEP Value
0.60th Quantile	1.0686
0.65th Quantile	0.3047
0.70th Quantile	0.0852
0.75th Quantile	2.6781
0.80th Quantile	3.8323

Table 1 shows that the principal component quantile regression model with the 0.70th quantile had the smallest RMSEP value, which was 0.0852, so it can be concluded that the model is good at predicting non-invasive blood glucose levels. The best quantile is also influenced by the distribution of data. Figure 4 shows that the data overfilled above the median (sloping right). This is one of the best quantile selection factors in the model.

IV. CONCLUSION

The result of the study showed that the characteristic data of a good quantile regression calibration model for predicting non-invasive blood glucose levels which were based on the summary data on period immensity and the principal component analysis (95% cumulative diversity proportion) was a model with a quantile of 0.70.

References

[1] A Djuraidah, A.H. Wigena, “Regresi kuantil untuk Eksplorasi Pola Curah Hujan di Kabupaten Indramayu,”Jurnal Ilmu Dasar, 2011,Vol. 12(1), pp. 50-56.

[2] A Kania, “Pendugaan Kadar Glukosa Darah Non-Invasif menggunakan Regresi Kuadrat Terkeil Parsial dengan Beberapa Pendekatan Peringkasan,” *Skripsi*, 2020.

[3] International Diabetes Federation, “IDF Diabetes atlas ninth edition”, 2019, ISBN 978-930229-87-4.

[4] Khikmah L, “Modeling Governance KB with CATPCA to Overcome Multicollinearity in the Logistic Regression,” di dalam: Khikmah L, Wijayanto H, Syafitri U D, editor. The 3rd International Conference on Mathematics, Science and Education, 2017, J. Phys, Ser. 824012027, IOP Publishing.

[5] Naes T, Issakson T, Fearn T, Davies T, *Multivariate Calibration and Classification*. United Kingdom: NIR Publications Chichester, 2002.

[6] Rosni, “Perbandingan Metode Regresi Komponen Utama Regresi Kuadrat Terkecil Parsial dan Support Vector Regression dalam Menduga Kadar Glukosa Darah Non-Invasif,” *Thesis*, 2019.

[7] Suri NNRR, Murty M N, Athitan G, *Outlier detection: techniques and applications*. Switzerland: Springer, 2019.

[8] Tonah, “Pemodelan Kalibrasi Peubah Ganda dengan Pendekatan Regresi Sinyal P-spline,” *Thesis*, 2006.

[9] World Health Organization, *Guidelines For The Prevention Management and Care of Diabetes Melitus*. Khatib Oussama M.N, editor. EMRO Technical Publications Series:32, 2006.

Authors

First Author – Siti Arita Novia,S.Si., college student, Department of Statistics, Faculty of Mathematics and Natural Science (FMIPA), IPB University, Bogor, 16680, Indonesia, and siti_arita@apps.ipb.ac.id
Second Author – Dr. Ir. Erfiani,M.Si., lecturer,Department of Statistics, Faculty of Mathematics and Natural Science (FMIPA), IPB University, Bogor, 16680, Indonesia, and erfiani@apps.ipb.ac.id.
Third Author – Prof. Dr. Ir, Aji HamimWigena,M.Sc., lecturer,Department of Statistics, Faculty of Mathematics and Natural Science (FMIPA), IPB University, Bogor, 16680, Indonesia, and aji_hw@apps.ipb.ac.id.

Correspondence Author – Prof. Dr. Ir, AjiHamimWigena,M.Sc., lecturer,Department of Statistics, Faculty of Mathematics and Natural Science (FMIPA), IPB University, Bogor, 16680, Indonesia, and aji_hw@apps.ipb.ac.id.