Modeling of hypertension risk factors using local linear of additive nonparametric logistic regression

E Ana¹, N Chamidah¹*, P Andriani², and B Lestari³

¹Department of Mathematics, Faculty of Science and Technology, Airlangga University. Jalan Mulyorejo Kampus C UNAIR, Surabaya 60115, Indonesia.
²Undergraduate Student of Statistical Study Program, Department of Mathematics, Faculty of Science and Technology, Airlangga University. Jalan Mulyorejo Kampus C UNAIR, Surabaya 60115, Indonesia.
³Faculty of Mathematics and Natural Sciences, The University of Jember. 37 Kalimantan Street, Campus of Tegal Boto, Jember 68121, Indonesia.

*Corresponding author’s e-mail: nur-c@fst.unair.ac.id

Abstract. Hypertension has become a serious health problem in Indonesia because of its prevalence, however, the causative factors could not be ascertained for about ninety percent of the patients. Various studies have found several risk factors causing hypertension to be obesity, family history, stress levels, heart rate, and an unhealthy lifestyle. In this case, the variables are considered influential on hypertension through a regression curve without a specific pattern. Also, we need to describe the functional relationships between several predictor variables with binary or dichotomous response variables and need to describe locally effect of predictor variables to the response variable. Therefore, in this study, to model the case of hypertension by age, body mass index, heart rate, stress levels, we use the additive nonparametric logistic regression approach based on local linear estimators. The results of the study showed that hypertension was most prevalent among respondents over 65 years of age with BMI between 25-30 kg/m² (obesity) and normal heart rate (60-100) bpm and most of them were found to be experiencing mild stress conditions. The model obtained a classification accuracy of 95 percent (in-sample) and 89.47 percent (out-sample) with a cut off probability value of 0.4.

1. Introduction
Hypertension is often referred to as the silent killer because it takes the life of affected individuals without showing symptoms. However, the factors causing the disease (around 90%) are still unknown. The number of people living with hypertension is predicted to become 1.56 billion worldwide by the year 2025 [1]. The sickness is associated with cardiovascular diseases (CVD) risk factors, incidence, and mortality [2]. It is also found to be prevalent among people of 35 years of age and above, currently smoking, and obese [3]. The Seventh Report on the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure created a category called "pre-hypertension" which was defined as a systolic blood pressure (SBP) of 120-139 millimeters of mercury (mmHg) and a diastolic blood pressure (DBP) of 80-89 mmHg [1]. Pre-hypertension, even in the low range (SBP: 120-130 mmHg or DBP: 80-85 mmHg), has been confirmed to have a higher risk of developing into hypertension [4].
Hypertension has been associated with increased risk of coronary artery and cardiovascular and cerebrovascular diseases [5-6]. A meta-analysis also reported that lower blood pressure could also lead to cardiovascular and chronic kidney diseases [7-8]. This situation is critical in the Southeast Asian region with studies reporting HTN as an important risk factor for the attributable burden [9-10]. Several studies have found different risk factors for hypertension such as obesity, family history, stress levels, heart rate, and an unhealthy lifestyle. Furthermore, previous research showed that the classification accuracy using binary logistic regression was to be 72.5352 percent greater than the C4.5 algorithm which was 64.0845 percent. Therefore, it could be said that binary logistic regression is better than the C4.5 algorithm [11].

However, the variables used were considered influential on hypertension through a regression curve without a pattern, therefore, a nonparametric approach was implemented. Moreover, logistic regression is a mathematical model used in describing the functional relationships between several predictor variables with binary or dichotomous response variables [12]. Binary response (Y) variables consist of two categories which are success (Y = 1) and failure (Y = 0), with respect to the Bernoulli distribution and the predictor variable is assumed to be linear to the parameters [13]. Generalized additive model (GAM) is an approach used in the exploration of nonparametric data [14].

Estimates for the approaches could be conducted by using smoothing techniques [15-22] and a local linear estimator is mostly used because of its ability to estimate functions locally at each point, thereby making the model obtained closer to the actual data pattern [23]. It is also useful because it can model the hypertension by age, body mass index, heart rate, and stress levels effectively due to its mean square error (MSE) value, which is close to zero, making the estimation results close to the observation value, and its determination coefficient which approaches 100 percent. This means that the regression model could describe the variability of risk factors for hypertension close to 100 percent [24]. In addition, the use of local linear in various cases of study for instance in medical and education cases study, have been introduced by researchers who have used local linear in [25-29].

Based on the facts from previous research, the hypertension risk factors of Surabaya Hajj Hospital based on additive nonparametric logistic regression approaches were modeled through the use of local linear estimator. The results of this research could be used to predict the chance of an individual suffering from hypertension showing the risk factors in order to reduce the prevalence of the disease in Indonesia.

2. Literature review

In this section we review some statistical concepts that can be used to support in modeling of hypertension risk factors.

2.1. Nonparametric Regression model

Nonparametric regression model is a model used in determining the relationship pattern between the response variable (y) and the predictor variable (t) where the function of the regression curve or the pattern of the two variables is unknown. The model generally follows the equation below:

\[ y_i = m(t_i) + \epsilon_i , \quad i = 1, 2, ..., n \]

where \( \epsilon_i \) is random error assumed as independent with zero mean and variance \( \sigma^2 \), and \( m(t_i) \) is the regression function which is to be estimated [30].

2.2. Logistic Regression model

Logistic regression is a mathematical model used in describing the functional relationships between several predictor variables with binary or dichotomous response variables [12]. These response (Y) variables consist of two categories which are success (Y = 1) and fail (Y = 0), with respect to the Bernoulli distribution and the predictor variable is assumed to be linear to the parameters [13]. The model is known as the logit model because it uses logit transformation as a link function from the following equation:
2.3. Additives nonparametric logistic Regression

The additive model is an illustration of the response variable which depends on the sum of several functions of several predictor variables. It could be expressed as follows:

\[ y_i = \sum_{j=1}^{p} m_j(t_{ij}) + \epsilon_i, \quad i = 1, 2, \ldots, n \]  

(3)

Next, by using the generalized additive model (GAM) approach, the equation (3) becomes:

\[ \log(\mu_i/(1-\mu_i)) = \sum_{j=1}^{p} m_j(t_{ij}) \]  

(4)

where

\[ \mu_i = \frac{e^{\sum_{j=1}^{p} m_j(t_{ij})}}{1 + e^{\sum_{j=1}^{p} m_j(t_{ij})}} \]  

(5)

The general algorithm for obtaining the estimated value of regression functions in the generalized additive model (GAM) is the local scoring algorithm.

2.4. Local linear estimator

The regression function could be estimated by a local linear estimator, through the equation below:

\[ \hat{m}(t) = X(t_0)\hat{\beta}(t_0) \]  

(6)

where \( \beta(t_0) \) is parameter of the model at target point \( t_0 \) that is estimated by using weighted least square (WLS) method, i.e., by minimizing the following function:

\[ Q(t_0) = (y - X(t_0)\beta(t_0))^T K_h(t_0)(y - X(t_0)\beta(t_0)) \]

such that we obtain:

\[ \hat{\beta}(t_0) = (X^T(t_0)K_h(t_0)X(t_0))^{-1}X^T(t_0)K_h(t_0)y \]  

(7)

Based on equations (6) and (7), the local linear estimator \( \hat{m}(t) \) could be written as:

\[ \hat{m}(t) = (t_0)^T X^T(t_0)K_h(t_0)X(t_0)^T K_h(t_0)y \]  

(8)

2.5. Selection of optimal bandwidth

Bandwidth is the control of the balance between the smoothness of the function and its goodness of fit to the data [30]. The best estimation of regression function depends on the optimal bandwidth \( h_{opt} \). The optimal bandwidth can be obtained by getting the solution of the following cross validation (CV) optimization [31]:

\[ CV(h) = \min \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}_{\alpha,i}(t_i))^2 \]  

(9)

2.6. Deviance

Deviance is a statistical value used to find the suitability of the model (goodness of fit) by comparing the actual to the expected model. Suppose that we have data set \( \{y_i\}, \quad i = 1, 2, \ldots, n \) which is binomially distributed \( BIN(1, \mu_i) \). The deviance \( D(y_i, \mu_i) \) could be calculated as follows [32]:

\[ D(y_i, \mu_i) = -2 \sum_{i=1}^{n} \left( y_i \log \mu_i + (1 - y_i) \log (1 - \mu_i) \right) \]  

(10)
2.7. Accuracy of classification

The accuracy of the classification model is calculated by using the apparent error rate (APER). The value obtained is used in classifying an object error. The APER is given by:

$$\text{APER} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\%$$

(11)

where $n_{rs}$ ($r = s = 1, 2$) is the number of observation in cell $rs$ that is intersection between row $r$ and column $s$. Therefore, the accuracy of classification is given by:

Accuracy of Classification = 100% - APER

(12)

3. Methods

The primary data used was obtained through questionnaires and taking of measurements on outpatients of the Heart Poly Surabaya Hajj Hospital. The research made use of 59 respondents consisting of 28 suffering from hypertension and 31 from other ailments with 40 in-sample and 19 out-sample data. The variables in the model used in this study are given in Table 1.

| Variable | Description | Scale |
|----------|-------------|-------|
| $Y$      | Hypertension incidents | Nominal |
|          | 0 = not suffering hypertension | |
|          | 1 = suffering hypertension | |
| $t_1$    | Age (Year) | Ratio |
| $t_2$    | Body Mass Index (kg/m$^2$) | Ratio |
| $t_3$    | Heart rate (bpm) | Ratio |
| $t_4$    | Stress level | Ratio |

The following steps were taken in analyzing the data:

- Descriptive statistical analysis was conducted on the predictor variables associated with the incidence of hypertension.
- Hypertension risk factors were modeled in Cardiac Poly outpatients in Surabaya Hajj Hospital with an additive nonparametric logistic regression approach based on the local linear local estimator with the following steps:
  - Determining of the kernel functions, i.e., Gaussian kernel functions;
  - Determining of the optimal bandwidth value based on minimum CV on each predictor;
  - Determining of the initial estimation value by using the optimal bandwidth value obtained from step (b);
  - Determining of the estimated value by using local scoring;
  - Testing of suitability of the model;
  - Analyzing and interpreting of the model of hypertension events.

4. Results

In this section we give the results of this study, i.e., description of variables, and modeling of hypertension risk factors.

4.1. Description of variables associated with hypertensive events

The characteristics of hypertension incidence are as shown in Table 2.
Table 2. Percentage of respondents’ characteristics of hypertension events in outpatients in heart poly Surabaya Hajj hospital.

| Variable         | Hypertension Incidents (%) |
|------------------|----------------------------|
|                  | Hypertension | Non-Hypertension |
| Age (Year)       |              |                 |
| < 45             | 7.14         | 61.3            |
| 45-54            | 10.71        | 22.58           |
| 55-64            | 35.71        | 12.9            |
| > 65             | 46.43        | 0               |
| < 25             | 35.71        | 70.97           |
| BMI (kg/m²)      |              |                 |
| 25-30            | 46.43        | 12.9            |
| > 30             | 17.86        | 16.13           |
| < 60             | 7.14         | 3.23            |
| Heart Rate (bpm) |              |                 |
| 60-100           | 85.71        | 77.42           |
| > 100            | 7.14         | 19.35           |
| Stress Level (total score) | 60.71 | 64.52 |
| 11-16            | 25           | 22.58           |
| ≥17              | 14.29        | 12.90           |

4.2. Modelling of hypertension risk factors
The selection of optimal bandwidth based on the minimum CV value was the first step conducted to get the appropriate estimation of hypertension incidence models. The bandwidth and CV for each predictor variable are given in Table 3.

Table 3. Bandwidth and CV for each predictor variable

| Predictor Variable | Optimal Bandwidth | Minimum CV |
|--------------------|-------------------|------------|
| Age                | 28.3              | 0.1448348  |
| BMI                | 8.6               | 0.2486765  |
| Heart rate         | 22.1              | 0.2448074  |
| Stress level       | 3.2               | 0.2790422  |

The results in Table 3 were used to determine the value of the initial estimation of the regression function for each predictor. Each of them was observed with a multi-predator local linear estimator, and a result of 77 years old, with BMI 26.84 kg/m², heart rate 84 bpm, and total stress level score 20 was found.

\[ \eta_i = 1.7657713 + 0.1357574(0) + 0.3868306 + 0.1160375(0) + \
0.789679 - 0.077819(0) + 1.5592301 + 0.1534729(0) \]

\[ = 4.50151156 \]

Therefore, the \( \hat{\mu}_i \) value is:

\[ \hat{\mu}_i = \frac{\exp(4.50151156)}{1 + \exp(4.50151156)} = 0.9890295 \]

Next, we test the suitability of the model used by calculating the deviance value which is given in (11). Hypothesis used for testing the suitability the model used is as follows:

\[ H_0 : \text{Model is appropriated} \quad \text{versus} \quad H_1 : \text{Model is not appropriated.} \]
Then, by applying Eq. (11) we get the deviance value $D(y_i, \mu_i) = 7.617064$ which is smaller than the value of the Chi-Square distribution, i.e., $\chi^2_{36,0.05} = 50.9985$. It means that the model used is in accordance with the nonparametric logistic regression model.

For the purpose of classification, a cut off probability was needed such that if the cut off the probability of $\hat{\mu}_i$ is greater, then the prediction result is 1 and if it is less, the prediction result is 0. Based on the output, the best cut off probability value obtained was 0.46. Therefore, the value of classification accuracy in the in-sample data is presented in Table 4.

| Observation                   | Prediction   | Total |
|-------------------------------|--------------|-------|
|                               | Without Hyp. | With Hyp. |   |
| Not suffering Hypertension    | 19           | 1      | 20 |
| Suffering Hypertension        | 1            | 19     | 20 |
| Total                         | 20           | 20     | 40 |

Based on Table 4, the APPER value could be estimated by applying equation (12) as follows:

$$\text{APPER} = \frac{1+1}{40} \times 100\% = 5\%.$$ 

So, we get the Classification Accuracy = 100% - 5% = 95%.

This shows that the estimation of the nonparametric logistic regression model based on the local linear estimator has the ability to explain 95 percent of hypertension incidence in the data in-sample.

Furthermore, 19 data out-sample were also estimated by using the same model used for the in-sample with a cut off probability value of 0.46 and a result of 67 years, with BMI 48.61 kg/m2, heart rate 79 bpm, and stress level scores of 10 were obtained. However,

$$\hat{\mu}_i = \frac{\exp(4.458179)}{1+\exp(4.458179)} = 0.988549202$$

The values of the accuracy of classification for the out-sample data is given in Table 5.

| Observation         | Prediction   | Total |
|---------------------|--------------|-------|
|                     | Without Hyp. | With Hyp. |   |
| Not suffering Hyp.  | 10           | 1      | 11 |
| Hypertension        | 1            | 7      | 8  |
| Total               | 11           | 8      | 19 |

The APPER value was obtained as follows:

$$\text{APPER} = \frac{1+1}{19} \times 100\% = 10.53\%$$

Classification Accuracy = 100% - 10.53% = 89.47%

This shows that the estimation of the nonparametric logistic regression model based on local linear estimators has the ability to explain 89.47% of hypertension incidence in the out-sample data.

5. Discussion

Hypertension has become a prevailing health problem in Indonesia because of its prevalence. However, it can be prevented by adequate control of a healthy lifestyle, routine blood pressure measurements, and other effective treatments.

Note that age, BMI, heart rate, and stress level are some of the factors observed to be triggering hypertension. As an example we select a 77 years old of person who represents the pre-very old
person. If a person is 77 years old, with a BMI of 26.84 kg/m², heart rate 84 bpm, and a stress level score of 20, such person has 98.9 percent risk of developing hypertension. This shows that an increase in those factors has a probability of causing the disease. The finding is in accordance with the results of previous studies that increasing age, BMI, smoking, diabetes, and additional salt intake are the common risk factors of hypertension [33-34].

6. Conclusion
Hypertension was found to be generally occurring in persons over the age of 65 years with a BMI between 25-30 kg/m² and a normal heart rate between 60-100 bpm with most of them experiencing mild stress. Since, the modelling conducted on in-sample data gave a classification accuracy of 95 percent and 89.47 percent for out-sample data with a cut off probability value of 0.46, then it could be concluded that the estimation model by using local linear estimator of additive nonparametric logistic regression is the best estimator for determining the hypertension incidence.

References
[1] Kearney P M, Whelton M, Reynolds K, Muntner P, Whelton P K, and He J 2005 Global Burden of Hypertension: Analysis of Worldwide Data Lancet 365 217–223.
[2] Elliot W J, and Black H R 2007 Prehypertension Nat Clin Pract Cardiovasc Med. 4(10) 538-548.
[3] Sathish T, Kannan S, Sarma P S, Razum O, Amanda G T, and Thankappan K R 2016 Risk Score to Predict Hypertension in Primary Care Settings in Rural India Asia-Pac J Public Health 28 26S-31S.
[4] Israeli E, Korzets Z, Tekes-Manova D, Tirosh A, Schochat T, Bernheim J, and Golan E 2007 Blood-pressure categories in adolescence predict development of hypertension in accordance with the European guidelines American journal of hypertension 20(6) 705-709.
[5] Qureshi A I, Suri M F, Kirmani J F, Divani A A, and Mohammad Y 2005 Is prehypertension a risk factor for cardiovascular diseases? Stroke 36(9) 1859-63.
[6] Wu S, Huang Z, Yang X, Li S, Zhao H, Ruan C, Wu Y, Xin A, Li K, Jin C, and Cai J 2013 Cardiovascular events in a pre-hypertensive Chinese population: four-year follow-up study International Journal of Cardiology 167(5) 2196-2199.
[7] Huang Y, Wang S, Cai X, Mai W, Hu Y, Tang H, and Xu D 2013 Pre-hypertension and incidence of cardiovascular disease: a meta-analysis BMC Medicine 11 177.
[8] Kim M J, Lim N K, and Park H Y 2012 Relationship between pre-hypertension and chronic kidney disease in middle-aged people in Korea: the Korean heart and epidemiology study BMC Public Health 12 960.
[9] Lim S S, Vos T, and Flaxman A D 2012, A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010 The Lancet 380(9859) 2224-2260.
[10] Lopez A D, Mathers C D, Ezzati M, Jamison D T, and Murray C J 2006 Global and Regional burden of disease and risk factors 2001: systematic analysis of population health data The Lancet 367(9524) 1747-1757.
[11] Rumaenda W, Wilandari Y, and Safitri D 2016 Perbandingan Klasifikasi Penyakit Hipertensi Menggunakan Regresi Logistik Biner dan Algoritma C4.5 Jurnal Gaussian 3(2) 299-309.
[12] Kleinbaum D G 1994 Logistic Regression Springer New York.
[13] Hosmer D, Lemeshow 2000 Applied Logistic Regression John Wiley and Sons Inc USA.
[14] Hastie T J and Tibshirani R J 1990 Generalized Additive Models Chapman & Hall London.
[15] Chamidah N and Lestari B 2016 Spline Estimator in Homoscedastic Multi-Response Nonparametric Regression Model in Case of Unbalanced Number of Observations Far East Journal of Mathematical Sciences (FJMS) 100 1433-1453.
[16] Lestari B, Anggraeni D, and Saifudin T 2017 Konstruksi dan Estimasi Matriks Kovariansi dalam Model Regresi Nonparametrik Multirespon Berdasarkan Estimator Smoothing Spline untuk Beberapa Kasus Ukuran Sampel Proceeding of National Seminar on Mathematics and Its Applications (SNMA) 2017 Airlangga University Surabaya pp. 238-242.

[17] Lestari B, Fatmawati and Budiantara I N 2017 Estimasi Fungsi Regresi Nonparametrik Multirespon Menggunakan Reproducing Kernel Hilbert Space Berdasarkan Estimator Smoothing Spline Proceeding of National Seminar on Mathematics and Its Applications (SNMA) 2017 Airlangga University Surabaya pp. 243-250.

[18] Lestari B, Anggraeni D and Saifudin T 2018 Estimation of Covariance Matrix based on Spline estimator in Homoscedastic Multi-Responses Nonparametric Regression Model in Case of Unbalance Number of Observations Far East J.Math.Sciences (FJMS) 108(2) 341-355.

[19] Islamiyati A, Fatmawati and Chamidah N 2018 Estimation of Covariance Matrix on Bi-Response Longitudinal Data Analysis with Penalized Spline Regression Journal of Physics: Conf. Series Vol. 979 pp. 012093.

[20] Lestari B, Fatmawati and Budiantara I N 2019 Spline Estimator and Its Asymptotic Properties in Multiresponse Nonparametric Regression Model Songklanakarin Journal of Science and Technology In Press.

[21] Lestari B, Fatmawati, Budiantara I N and Chamidah N 2018 Estimation of regression function in multi-response nonparametric regression model using smoothing spline and kernel estimators Journal of Physics: Conference Series 1097 012091.

[22] Lestari B, Chamidah N and Saifudin T 2019 Estimasi fungsi regresi dalam model regresi nonparametrik birespon menggunakan estimator smoothing spline dan estimator kernel Jurnal Matematika, Statistika and Komputasi (JMSK) 15(2) 20-24.

[23] Hardle W 1990 Applied Nonparametric Regression Cambridge University Press New York.

[24] Adiwati, T and Chamidah, N. 2019 Modelling of Hypertension Risk Factors Using Penalized Spline to Prevent Hypertension in Indonesia IOP Conf. Series: Materials Science and Engineering 546 052003 doi:10.1088/1757-899X/546/5/052003.

[25] Chamidah N, Tjahjono E, Fadilah A R., and Lestari B 2018 Standart Growth Charts for Weight of Children in East Java Using Local Linier Estimator Journal of Physics: Conference Series 1097 012092.

[26] Nidhomuddin, Chamidah N, Kurniawan A 2019 Admission Test Modelling of State Islamic College in Indonesia Using Local Linear for Bivariate Longitudinal Data IOP Conf. Series: Materials Science and Engineering 546 052047 doi:10.1088/1757-899X/546/5/052047.

[27] Darnah, Utoyo M I and Chamidah N 2019 Modeling of Maternal Mortality and Infant Mortality Cases in East Kalimantan using Poisson Regression Approach Based on Local Linear Estimator IOP Conf. Series: Earth and Environmental Science 243 012023 doi:10.1088/1755-1315/243/1/012023.

[28] Puspitawati A., and Chamidah N 2019 Choroidal Neovascularisation Classification on Fundus Retinal Images Using Local Linear Estimator IOP Conf. Series: Materials Science and Engineering 546 052056 doi:10.1088/1757-899X/546/5/052056.

[29] Chamidah N, Gusti K H, Tjahjono E and Lestari B 2019 Improving of classification accuracy of cyst and tumor using local polynomial estimator TELKOMNIKA, Vol.17, No.3, pp.1492-1500.

[30] Eubank R L 1998 Nonparametric Regression and Spline Smoothing 2nd Edition Marcel Dekker New York.

[31] Gyorfi L, Kohler M, Krzyzak A, and Walk H 2002 A Distribution-Free Theory of Nonparametric Regression Springer.

[32] Collett D 2002 Modelling Binary Data 2nd Edition Chapman & Hall/CRC.
[33] Devi P, Rao M, Sigamani A, Faruqui A, Jose M, Gupta R, Kerkar P, Jain R K, Joshi R, Chidambaram N, Rao D S, Thanikachalam S, Iyengar S S, Verghese K, Mohan V, Pais P, and Xavier D 2012 Prevalence, risk factors and awareness of hypertension in India: a systematic review Journal of Human Hypertension 27(5) 281-287.

[34] Anchala R, Kannuri N K, Pant H, Khan H, Franco O H, Angelantonio E D, and Prabhakaran D 2014 Hypertension in India: a systematic review and meta-analysis of prevalence, awareness, and control of hypertension Journal of Hypertension 32(6) 1170.