Modelling of Hypertension Risk Factors Using Penalized Spline to Prevent Hypertension in Indonesia

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052003 doi:10.1088/1757-899X/546/5/052003

Modelling of Hypertension Risk Factors Using Penalized Spline to Prevent Hypertension in Indonesia

Tati Adiwati[1] and Nur Chamidah[2]

[1]Student of Study Program of Statistics, Department of Mathematics, Faculty of Science and Technology, Airlangga University
[2]Department of Mathematics, Faculty of Science and Technology, Airlangga University
Corresponding author's email : nur-c@fst.unair.ac.id

Abstract. Hypertension is an increase in blood pressure that increases to a target organ, such as stroke, coronary heart disease, right ventricular hypertrophy. Hypertension occurs if the blood pressure reaches 140 mmHg or more and diastole reaches 90 mmHg or more. According to WHO, from 50% of hypertensive patients recovering, only 25% received treatment, and only 12.5% could be treated well. Nationally, 25.8% of Indonesia's population suffers from hypertension. In this study, we modeled the risk of hypertension by considering age, heart rate, family hypertension, stress levels, and the body's future index as factors that influence the risk of hypertension. The cross-sectional survey was conducted in August 2018 at the Surabaya Hajj Hospital. Based on previous research the method used is logit and gompit logistic regression method, but the results obtained are not maximal. Therefore, in this study the researchers proposed a method for constructing hypertension risk factor modeling using a nonparametric application using a penalized spline estimator. The result of classification accuracy by using non-parametrical is 96%. Based on the result, we conclude that non-parametrical approach has better than outcome so that it can be used to modelling the risk of hypertension.

1. Introduction

Hypertension or high blood pressure in a place where a person increases blood pressure for a long time which successfully improves mortality [1]. Hypertension is an increase in blood pressure that increases to a target organ, such as stroke, coronary heart disease, right ventricular hypertrophy [2]. The criteria for hypertension used in the determination of cases are the measurement results of systolic blood pressure ≥140 mmHg or diastolic blood pressure ≥90 mmHg [3]. According to WHO, from 50% of hypertensive patients recovering, only 25% received treatment, and only 12.5% could be treated well. Based on the Household Health Survey (SKRT) in 2004, the prevalence of hypertension in Java was 41.9%, with ranges in each province 36.6-47.7%. Urban prevalence is 39.9% and in rural areas 44.1% [4].

Hypertension is not a disease with a single factor, but is caused by many factors, namely obesity, unhealthy eating patterns, lack of physical activity, psychological stress conditions, alcohol drinking habits, coffee consumption patterns and smoking habits [5]. Decreasing hypertension in people with risk factors for the type, age above 18 years, has a family history of hypertension, and in people who smoke [6]. In a previous study using the Prevalence Odd Ratio (POR) Method stating that someone

has obesity by 3.8 times can suffer from hypertension, and someone who has a risk of having a risk of 6.2 times can suffer from hypertension [7].

Previous research on hypertension has been carried out by several researchers. [8] using obtaining computational gompit and logit regression and getting the results of classification accuracy of 81.5% and 85.2%. The data used in this study are categorized into two categories: success and failure. One of the right methods is used to model the probability of someone issuing hypertension or not is a nonparametric regression with a spline punished estimator.

Splines are pieces of polynomials that have different segments, which are combined together at vertices [9]. It is this segmented nature that gives more than ordinary polynomials to adapt effectively to the local characteristics of the function or data [10]. The model estimator form punishes the spline multipredictor using an additive model with the response variable (y) needed for the sum of the predictor variables (x). Spline estimator has been studied by [11-13]. With this study, it is expected that the incidence of hypertension in Indonesia can reduce the risk of hypertension.

2. Research Methods

The data used in this study are primary data obtained from questionnaires and interviews with Cardiac Poly in Surabaya Haji Hospital which was conducted from August to September 2018 with 54 respondents. The variables used for this study consist of response variables and predictor variables. The response variable (Y) is the incidence of hypertension, while the predictor variables used are age (X_1), body mass index (X_2), heart rate (X_3), and stress (X_4). The research step is the steps that must be taken to solve existing problems. The steps to identifying using nonparametric logistic regression approach using OSS-R based on these following steps:

1) estimation of f_i for each predictor with the following steps :
   a. determine the order of the polynomial, number of knots, and the smoothing parameters (λ) based on the minimum GCV value
   b. defines the X_j matrix by entering the optimal polynomial order and knots point
   c. define the estimation value of $\hat{\beta}_j$ by entering the optimal smoothing parameter value
   d. calculate the $\hat{f}_j(X_j) = X_j(X_j{}^T X_j + n\lambda_j D_j)^{-1} X_j{}^T Y$
2) iteration of local scoring and backfitting algorithms to obtain an additive model based on the penalized spline estimator with the following steps :
   a. defines the response variable and predictor variable
   b. determine the initial value to be used in the iteration to 0 $(h = 0)$
   c. determine local scoring for $(h = 0,1,2, \dots)$ with the following steps :
      i. determine the partial residual $R_j{}^{(h+1)} = z - \sum_{s=1}^{j-1} f_s{}^{(h)}(X_s) - \sum_{s=j+1}^{p} f_s{}^{(h)}(X_s)$
      ii. determine the smoothing function $f_j{}^{(h+1)} = H(\lambda_j) R_j{}^{(h+1)}$
      iii. determine the RSS value $RSS^{(h+1)} = \frac{1}{n}\{(y - \hat{\mu})^T (y - \hat{\mu})\}$
      iv. iterates steps (1) to (3) to obtain RSS that meets convergent criteria $\left|RSS^{(h+1)} - RSS^{(h)}\right| < \varepsilon$
      v. determine vector adjusted dependent variable $z_i{}^{(h+1)}$, $\mu_i{}^{(h+1)}$, and $\eta_i{}^{(h+1)}$
      vi. determine the $W_i{}^{(h+1)}$ matrix
      vii. calculate $avg(Dev) = \frac{1}{n}\{(y_i - \mu_i)^T W_i (y_i - \mu_i)\}$
      viii. iterates steps (1) to (7) to obtain $avg(Dev)$ that meets the convergent criteria $\left|avg(Dev)^{(h+1)} - avg(Dev)^{(h)}\right| < \varepsilon$
3) analyze the results of classification accuracy with the following steps:

a. describe the cut off probability value as the dividing boundary between categories 0 and 1 in classifying objects
b. calculate estimate of Y value
c. get the best cut off probability value
d. calculate the APPER value $APPER = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \times 100\%$
e. calculate $classification\ accuracy = 100 - APPER$
f. calculate $Press'Q = \frac{[N-(nK)]^2}{N\ (K-1)}$ and compare the value of $Press'Q$ with the value of $ChiSquare$ with free degree 1

3. Result and Analysis

The first step to estimate the nonparametric regression model based on the penalized spline estimator is to determine the order, many knot points, knot points, and optimal smoothing parameters for each predictor based on the minimum GCV criteria. We obtain the result for each predictor as given in Table 1.

Table 1 Optimum Lambda Value for each Predictor Variable

No	Predictor Variable	Orde	Total Knot	Knot Point	Minimum GCV	Optimum Lambda
1	X1	1	1	50	0.1202796	12
2	X2	1	2	22.22222 26.23518	0.2440173	2.4
3	X3	1	2	76 91	0.2405815	2.9
4	X4	1	1	9.5	0.2671597	9999.9

After obtaining the optimal initial value for each predictor, the next step is to iterate using the local scoring algorithm, with estimated parameter values are as follow:

$\hat{\beta}_1 = [-27.15279483 \quad 0.49785308 \quad -0.05616246]^T$

$\hat{\beta}_2 = [-21.0895797 \quad 0.8481306 \quad 0.1088974 \quad -0.5527757]^T$

$\hat{\beta}_3 = [13.3830106 \quad -0.1455019 \quad -0.3671064 \quad 0.7035412]^T$

$\hat{\beta}_4 = [-3.7884923004 \quad 0.5079612398 \quad 0.0001308514]^T$

The form of penalized spline estimator for the initial value of function for each predictor in the first observation is as follows:

a. $\hat{f}_1(X_{1i}) = -27.15279483 + 0.497853086X_1 - 0.05616246(X_1 - 50)_+$

with terms:

$$\hat{f}_1(X_{1i}) = \begin{cases} -27.15279483 + 0.497853086X_1 & ; X_1 < 50 \\ -27.15279483 + 0.497853086X_1 - 0.05616246(X_1 - 50) & ; X_1 \geq 50 \end{cases}$$

b. $\hat{f}_2(X_{2i}) = -21.0895797 + 0.8481306X_2 + 0.1088974(X_2 - 22.22222)_+ - 0.5527757(X_2 - 26.23518)_+$

with terms:

$$\hat{f}_2(X_{2i}) = \begin{cases} -21.0895797 + 0.8481306X_2 & ; X_2 < 22.22222 \\ -23.5095217 + 0.957028X_2 & ; 22.22222 \leq X_2 < 26.23518 \\ -9.00735171 + 0.4042523X_2 & ; X_2 \geq 26.23518 \end{cases}$$

c. $\hat{f}_3(X_{3i}) = 13.3830106 - 0.1455019X_3 - 0.3671064(X_3 - 76) + 0.7035412(X_3 - 91)_+$

with terms:

$$\hat{f}_3(X_{3i}) = \begin{cases} 13.3830106 - 0.1455019X_3 & ; X_3 < 76 \\ 41.283097 - 0.5126083X_3 & 76 \leq X_3 < 91 \\ -22{,}7391522 + 0.1909329 & ; X_3 \geq 91 \end{cases}$$

d. $\hat{f}_4(X_{4i}) = -3.7884923004 + 0.5079612398X_4 + 0.0001308514(X_4 - 9.5)_+$

with terms:

$$\hat{f}_4(X_{4i}) = \begin{cases} -3.7884923004 + 0.5079612398X_4 & ; X_4 < 9.5 \\ -3.788361449 + 0.5067181515X_4 & ; X_4 \geq 9.5 \end{cases}$$

Each observation on the penalized spline estimator has a model, so that in this thesis an explanation will be given for the 34th observation only, for other observations the process carried out is the same as the 34th observation. The estimated results of additive nonparametric logistic regression models based on the penalized spline estimator for the 30th observation are as follow:

a. The first predictor variable at 34th observation is equal to 25. This value is include in criteria $X_1 < 50$, so the function used for $(X_{1,34})$ is
$\hat{f}_1(X_{1,34}) = -27.15279483 + 0.497853086X_1 \quad ; X_1 < 50$
$\hat{f}_1(X_{1,34}) = -27.15279483 + 0.497853086(25)$
$= -14.70646768$

b. The second predictor variable at 34th observation is equal to 19.22768787. This value is include in criteria $X_2 < 22.22222$, so the function used for $(X_{2,34})$ is
$\hat{f}_2(X_{2,34}) = -21.0895797 + 0.8481306X_2$
$\hat{f}_2(X_{2,34}) = -21.0895797 + 0.8481306(19.22768787)$
$= -4.781989250204178$

c. The third predictor variable at 34th observation is equal to 83. This value is include in criteria $76 \leq X_3 < 91$, so the function used for $(X_{3,34})$ is
$\hat{f}_3(X_{3,34}) = 41.283097 - 0.5126083X_3 \quad ; 76 \leq X_3 < 91$
$\hat{f}_3(X_{3,34}) = 41.283097 - 0.5126083(83)$
$= -1.2633919$

d. The fourth predictor variable at 34th observation is equal to 16. This value is include in criteria $X_4 \geq 9.5$, so the function used for $(X_{4,34})$ is

$$\hat{f}_4(X_{4,34}) = -3.788361449 + 0.5067181515X_4$$
$$\hat{f}_4(X_{4,34}) = -3.788361449 + 0.5067181515(16)$$
$$= 4.3189981236$$

In complete terms, the penalized spline estimator for the 34th observation model is:

$$\eta_{34} = \sum_{j=1}^{4} \hat{f}_j(X_{j,34})$$
$$= \widehat{f}_1(X_{1,34}) + \hat{f}_2(X_{2,34}) + \hat{f}_3(X_{3,34}) + \hat{f}_4(X_{4,34})$$
$$= -14.70646768 - 4.781989250204178 - 1.2633919 + 4.3189981236$$
$$= -16.43285070660418$$

The next step is to calculate the estimated value for the 34th observation

$$\hat{\mu}_{34} = \frac{\exp(-16.43285070660418)}{1 + \exp(-16.43285070660418)} = 0.00000007299676$$

Then, we classify categories $Y = 0$ for hypertension, and $Y = 1$ for normal. This is done by determining the threshold value is used as a comparison or cut off in the identification of hypertension contained in Table 2**.**

Table 2. Threshold Values and Accuracy Classification

No	Threshold	Accuracy Classification	No	Threshold	Accuracy Classification
1	0.00	76.27119	9	0.91	93.22034
2	0.03	86.44068	10	0.94	91.52542
3	0.07	88.13559	11	0.95	86.44068
4	0.09	91.52542	12	0.96	84.74576
5	0.11	93.22034	13	0.97	81.35593
6	0.52	94.91525	14	0.98	79.66102
7	0.64	96.61017	15	0.99	76.27119
8	0.88	94.91525	16	1.00	52.54237

Threshold that will be used as reference for cut-off category 0 or category 1 is determined by looking at the highest classification accuracy score and highest threshold value which has the highest classification accuracy. If value $\widehat{m}_i$ is greater than threshold value, then it will be classified as normal, vice versa. The classification result based on value $\widehat{m}_i$ in the 34th observation approaches are given in Table 3.

Table 3 Classification Accuracy

Observation	Prediction		Total
	Hypertension	Normal	
Hypertension	29	2	31
Normal	0	28	28
Total	27	32	59

obtained APPER value which is an opportunity for errors in classifying objects as follows

$$APPER = \frac{0+4}{27+0+4+28} \times 100 = 3.39\ \%$$

Based on the calculation obtained the value of classification accuracy of 96.6%, so it can be seen that the estimation of nonparametric regression models based on the penalized spline estimator obtained is valid for calculating the incidence of hypertension. The last step Press 'Q test needs to be done to determine the stability in the classification accuracy of the extent to which groups can be separated using existing variables by comparing the Press' Q value with the Chi-Square table value with a free degree 1. We validate the classification accuracy of nonparametric regression model approach with Press'Q value as follow:

$$Press'Q = \frac{(N-(nK))^2}{N(K-1)} = \frac{(59-((57)(2)))^2}{59(2-1)} = 51.2712$$

The Press'Q is compared with $\chi^2_{(0.05,1)} = 3.841$. Because of the Press'Q = 51.2712 is greater than $\chi^2_{(0.05,1)} = 3.841$, so it can be concluded that the model is stable or consistent.

But based on [8] research about hypertension risk factors too, obtained logit and gompit models are $g(x) = -2,004 + 0,251X_1 - 0,158X_2 - 0,931X_{3(1)} + 0,144X_{4(1)} - 0,442X_{5(1)} + 0,277X_6$ and $g(x) = -1,329 + 0,179X_1 - 0,117X_2 - 0,885X_{3(1)} - 0,074X_{4(1)} - 0,162X_{5(1)} + 0,183X_6$

the research explained that the models is significant, but there is no test and information that the models is stable or consistent. Then on APPER value results that logit and gompit's classification accuracy is smaller than penalized spline's, it means 85.2% and 81.5%.

4. Conclusion

Modeling of hypertension risk factors by using logistic nonparametric regression based on penalized spline estimator is better than logistic parametric regression by using link function gompit and logit that it can increase the accuracy of the classification of hypertension risk factors up to 96%. It is expected that the public can realize the importance of a healthy lifestyle in order to maintain health in order to control blood pressure with, so as to avoid hypertension and as an effort to reduce the incidence of hypertension in Indonesia.

References

[1] Yeyeh. R. 2010 Asuhan Kebidanan 4 Patologi *Trans Info Media*
[2] Wikansari, N., Kertia, N., and Dewi, F. S. T. 2017 Determinants of smoking cessation

behaviour in people with hypertension *BKM Journal of Community Medicine and Public Health* 33(3) pp 135-140

[3] Badan Penelitian dan Pengembangan Kesehatan Kementrian Kesehatan RI 2013, Riset Kesehatan Dasar RISKESDAS 013 *Badan Litbangkes* Jakarta

[4] Sulistiyono, H., and Isnawati, M. 2011 Pemberian jus belimbing Demak (Averrhoe carambola l) berpengaruh terhadap penurunan tekanan darah sistolik dan diastolik pada penderita hipertensi *JURNAL GIZI KLINIK INDONESIA* 7(3) pp 123-128

[5] Dhianningtyas, Y., and Hendrati, L. Y. 2006 Risiko Obesitas, kebiasaan merokok, dan konsumsi garam terhadap kejadian hipertensi pada usia produktif *The Indonesian Journal of Public Health* 2(3) pp 105-109

[6] Departemen Kesehatan RI 2006, Pedoman Teknis Penemuan dan Tata Laksana Penyakit Hipertensi

[7] Korneliani, K., and Meida, D. 2015 Obesity and Stress with Hypertension *Journal of Public Health* 7(2) pp 117-121

[8] Andriani, P., and Chamidah, N. 2018 Modelling of Hypertension Risk Factors Using Logistic Regression to Prevent Hypertension in Indonesia *International Conference of Mathematics* in press

[9] Eubank, R. 1998 Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York

[10] Ruppert, D. 2002 Selecting The Number of Knots for Penalized Spline *Journal of Computational and Graphical Statistics* 11(4) 735-757

[11] 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**(9) 1433-1453

[12] 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

[13] 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, 979 012093, IOP Publishing. doi:10.1088/1742-6596/979/1/012093