Modelling of Hypertension Risk Factors Using Logistic Regression to Prevent Hypertension in Indonesia

Putri Andriani\textsuperscript{1} and Nur Chamidah\textsuperscript{2,*}
\textsuperscript{1}Student of Study Program of Statistics, Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia.
\textsuperscript{2}Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia.

\textsuperscript{*}Corresponding author email: nur-c@fst.unair.ac.id

Abstract. Hypertension called as the silent killer, is the number one non-infectious disease that causes death in the world every year. There are 185,857 cases recorded in 2018 in Indonesia. In this study, we model the hypertension risk by considering age, heart rate, hypertension history of family, eating salty foods, and smoking or exposure to cigarette smoke as the influence factors of hypertension risk. A cross-sectional survey was conducted in August 2018 at the Haji Hospital of Surabaya. Logistic regression is used to analyse the influence of various risk factors on hypertension and non-hypertension. In addition, we compare between logit and gompit link functions in logistic regression to build the modelling of hypertension risk factors based on the accuracy of the classification model. By using logit and gompit link functions, we obtain percentage of the classification accuracy are 85.2\% and 81.5\%, respectively. It means that the logit link function is better than the gompit link function for modelling hypertension risk factors. For these link functions, the significant factors that influence hypertension are age and heart rate.

Keywords: hypertension, risk factors, logistic regression, logit and gompit link functions.

1. Introduction
Hypertension called as the silent killer, is the number one non-infectious disease that causes death in the world every year [1]. Hypertension or high blood pressure is an increase in abnormal blood pressure in the arteries that carry blood from the heart to the entire body continuously for more than one period [2]. The prevalence of hypertension will continue to increase sharply and is predicted in 2025 as many as 29\% of adults worldwide develop hypertension. Hypertension has resulted in the deaths of around 8 million people every year, of which 1.5 million deaths occur in Southeast Asia, where 1 per 3 of the population has hypertension [3]. There were 185,857 cases recorded in 2016 in Indonesia [4].

Hypertension is still a health problem in Indonesia because every year the number of hypertension cases increases. 90\% is not known with certainty the cause of hypertension, but from various studies have found several risk factors that cause hypertension, namely obesity, family history, and unhealthy lifestyles such as consuming alcohol, excessive salt intake, smoking, and lack of physical activity [5]. In addition, stress is also a risk factor for hypertension [6]. Obesity and stress are factors that significantly influence the incidence of hypertension. In the previous study using the Prevalence Odd Ratio (POR) method stated that someone who is obese has a risk of 3.8 times suffering from
hypertension, and someone who is stressed has a risk of 6.2 times suffering from hypertension [7].

Previous research on hypertension has not been able to model the probability of someone developing hypertension. The right method is used to model the chance of an individual experiencing hypertension or not is logistic regression.

Logistic regression model is a mathematical model used to describe the functional relationship between several predictor variables and binary or dichotomous response variables [8]. The binary response variable \( Y \) is divided into two categories, namely success \( Y = 1 \) and failure \( Y = 0 \), and predictor variables are assumed to be linear to the parameters. With the existence of this research, it is expected that the community can maintain health in order to be able to control blood pressure with a healthy lifestyle, by avoiding hypertension and as a preventive measure so that the incidence of hypertension in Indonesia decreases.

2. Literature Review

2.1 Logit Link Function

Logistic regression is a method of data analysis that is used to find the relationship between the response variable \( Y \) that is binary or dichotomous and polycotomous predictor variable \( X \) [8]. Binary logistic regression model with the outcome of the response variable \( y \) consists of two categories, namely “success” and “fail” denoted by \( y = 1 \) (success) and \( y = 0 \) (failed) then the \( Y \) response variable follows the Bernoulli distribution [8], with a probability function.

\[
\begin{align*}
    f(x_i) &= \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}, y_i = 0,1, \ldots, m_i \\
    \pi(x) &= \frac{e^{\mathbf{x}' \mathbf{\beta}}}{1 + e^{\mathbf{x}' \mathbf{\beta}}} \\
    g(x) &= \ln \left( \frac{\pi}{1 - \pi} \right)
\end{align*}
\]

Binary logit model is a regression model based on the concept of probability. A binary logit model is expressed in the form:

\[
g(x) = \mathbf{x} \mathbf{\beta}; i = 1,2,\ldots,N
\]

where \( g(x) = \ln \left( \frac{\pi}{1 - \pi} \right) \) is the logit link function, \( \mathbf{x}_i = (1, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ik_i}) \) is a vector of the \( i \)-th predictor variable, and \( \mathbf{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})' \) is a vector of parameters corresponding to the predictor variables.

Logistic regression models are formed by expressing the value of \( E(Y = 1|x) \) as \( \pi(x) \), where \( \pi(x) \) is given as follows

\[
\begin{align*}
    \pi(x) &= \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \\
    \pi(x) &= \frac{\exp(g(x))}{1 + \exp(g(x))}
\end{align*}
\]

where \( p \) is the number of predictor variables. Furthermore, equation (3) can be expressed as in the following equation:

\[
\begin{align*}
    \pi(x) &= \frac{\exp(g(x))}{1 + \exp(g(x))}
\end{align*}
\]

2.2 Gompit Link Function

Binary gompit model is a regression model with a response variable consisting of two categories based on Gompertz distribution [10]. Logistic regression model with the gompit link function is given as follows:

\[
g(x) = \ln(-\ln(1 - \pi(x)))
\]

Based on equation (5), we obtain probability of success at \( i \)-th trial of predictor variables corresponding to \( X_i \), that is:

\[
\pi(x) = 1 - \exp[-\exp(\mathbf{x} \mathbf{\beta})]
\]
2.3 Maximum Likelihood Estimation

The parameter estimation method used is the Maximum Likelihood Estimation (MLE) method. Based on equation (1), we obtain the following likelihood function:

\[ L(\beta) = \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N} \pi(x_i)^{y_i} (1 - \pi(x_i))^{n_i - y_i} \]  

From equation (7), the log-likelihood function is obtained as follows:

\[ \ell(\beta) = \sum_{i=1}^{N} \left[ y_i \ln \pi(x_i) + (n_i - y_i) \ln(1 - \pi(x_i)) \right] \]  

Next, parameter estimation \( \beta \) can be obtained by maximizing \( \ell(\beta) \). These terms is simply so that the function \( \ell(\beta) \) reaches the maximum value if \( \frac{\partial \ell(\beta)}{\partial \beta_j} = 0 \), i.e.,

\[ \frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} X_{ij} (y_i - m_i \pi_i) = 0; j = 0, 1, 2, \ldots, K - 1 \]  

Since the derivative of \( \ell(\beta) \) with respect to \( \beta \) is a nonlinear function of \( \beta \), then to estimate it we use Newton-Raphson method, so that we obtain the estimator of parameter \( \beta \), i.e., \( \hat{\beta} \).

2.4 Parameter Testing

This test is done to determine the influence of the predictor variables on the variable response. Simultaneous tests are used to test the effect of predictor variables on response variables [8], with the following hypothesis:

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0 \]
\[ H_1 : \text{There is at least one } \beta_j \neq 0; j = 1, 2, \ldots, K - 1 \]

The statistical test used is the likelihood ratio test as follows:

\[ G = -2 \ln \left[ \frac{\left( \frac{n_0}{n} \right)^{n_0} \left( \frac{n_1}{n} \right)^{n_1}}{\prod_{i=1}^{n} \left[ \hat{\pi}_i \right]^{y_i} \left[ 1 - \hat{\pi}_i \right]^{n_i - y_i}} \right] \]  

with \( n_0 = \sum_{i=1}^{n} (1 - y_i) \) and \( n_1 = \sum_{i=1}^{n} y_i \), where \( n_0 \) is the number of observation values \( Y = 0 \), \( n_1 \) is the number of observation values \( Y = 1 \), and \( n \) is total number of observations \( n = n_0 + n_1 \). The critical area of the hypothesis testing is that \( H_0 \) is rejected if \( G > \chi^2_{(\alpha, K-1)} \) or \( p-value < \alpha \).

If the test results are simultaneously significant, then we will do a partially testing for each predictor variable by using the following hypothesis:

\[ H_0 : \beta_j = 0 \]
\[ H_1 : \beta_j \neq 0; j = 1, 2, \ldots, K - 1 \]

The statistical test used is the Wald test as follows:

\[ Z_j = \frac{\hat{\beta}_j}{S(\hat{\beta}_j)}; j = 1, 2, \ldots, K - 1 \]  

The critical area of the hypothesis testing is that \( H_0 \) is rejected if \( |Z_j| > Z_{\alpha/2} \).

To find out the comparison between the probability of an event occurring and the probability of no occurring event, we use Odd Ratio (OR) as follows:
4

\[ OR = \frac{\pi(1)}{1 - \pi(1)} = \frac{\pi(1)[1 - \pi(0)]}{\pi(0)[1 - \pi(1)]} = e^\beta \]  \(12\)

The testing in (12) is done to conclude that the obtained model is appropriate. It means that there is no difference between the results of observations and possible predictions of the model.

Next, to test the suitability of the model, we use Hosmer and Lameshow statistical testing with the following hypothesis:

\[ H_0 : \text{The model is appropriate} \]
\[ H_1 : \text{The model is not appropriate} \]

The statistical test used is the Hosmer and Lameshow testing as follows:

\[ \hat{C} = \sum_{i=1}^{r} \left( \frac{o_i - c_i \hat{\mu}_i}{c_i \hat{\mu}_i(1 - \hat{\mu}_i)} \right)^2 \]  \(13\)

where \(o_i\) is the number of response variables in the group, \(\hat{\mu}_i\) is average estimated probabilities in the group, and \(c_i\) is the number of observations in the \(k\)-th group. The critical area of the hypothesis testing is that \(H_0\) is rejected if \(\hat{C} > \chi^2_{(r, \alpha)}\) or \(p\text{-value} < \alpha\).

2.5 Classification Accuracy

Classification procedure is an evaluation to see the possibility of misclassification carried out by a classification function by using the Apparent Error Rate (APPER) value which is the value of the wrong or incorrect proportion of the sample classified by the classification function [9]:

\[ APPER = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\% \]  \(14\)

where \(n_{11}\) is the number of failed events observations and classified failed from the results of predictions; \(n_{o_i}\) is the number of failed events observations and classified success from the results of predictions; \(n_{21}\) is the number of success events results the observations and classified failed from the results of predictions; \(n_{22}\) is the number of success events the observations and classified success from the results of prediction.

3. Research Method

The data used in this study is primary data obtained from questionnaires and interviews with Cardiac Poly at 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 \((X_1)\), heart rate \((X_2)\), family history of hypertension \((X_3)\), consumption of salty food \((X_4)\), smoking \((X_5)\), and stress \((X_6)\). The research steps are the steps that must be taken to solve existing problems.

In this study, the steps to compare between logit and gompit models are as follows:

1. Test the effect of predictor variables on response variables by using simultaneously testing of parameters for all predictor variables;
2. If the result in step (1) is significant, do the partially testing to determine the effect of each predictor variable on response variables;
3. Determine the logit model and gompit model from the results of analysis;
4. Test the suitability of the logit model and the gompit model that have been determined;
Determine the accuracy of classification on logit model and gompit model.

4. Result and Discussion

4.1 Test the Parameter Significance Simultaneously

The simultaneously parameter testing was used to determine the effect of suspected factors that affect hypertension simultaneously with the following hypothesis:

\[ H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \]
\[ H_1 : \text{There is at least one } \beta_j \neq 0; j = 1, 2, 3, 4, 5, 6 \]

Testing the parameter significance simultaneously can be seen from the G test statistic as given in Table 1.

| Link Function | G Test Statistic | p-value |
|---------------|------------------|---------|
| Logit         | 50.592           | 0.000   |
| Gompit        | 50.924           | 0.000   |

Table 1 shows that the G value for the Gompit model is larger than those for the logit model, so that the most significant \( \beta \) coefficient is the Gompit model. Simultaneous test result by using significant level \( \alpha = 5\% \) and \( p\text{-value} = 0.000 \) indicates that \( H_0 \) is rejected which means that there are one or more significant predictor variables parameters.

4.2 Test the Parameter Significance Partially

Parameter significance testing can be done partially to see variables that affect partially significant. Hypothesis used is as follows:

\[ H_0 : \beta_j = 0 \]
\[ H_1 : \beta_j \neq 0; j = 1, 2, 3, 4, 5, 6 \]

In Table 2, we give the \( \beta \) values that are obtained partially for Logit and Gompit models.

|            | Logit Model | Gompit Model |
|------------|-------------|--------------|
|            | \( \beta \) | Wald | p-value | \( \beta \) | Wald | p-value |
| Constant   | -2.00387    | -0.41 | 0.682   | -1.32940    | -0.40 | 0.692   |
| Age        | 0.25115     | 2.65  | 0.008*  | 0.17852     | 2.68  | 0.007*  |
| Heart Rate | -0.15779    | -2.31 | 0.021*  | -0.11730    | -2.43 | 0.015*  |
| Family History (1) | -0.93141 | -0.58 | 0.562   | 0.39        | -0.88512 | -0.83 | 0.404   |
| Consumption of Salty Food (1) | 0.14377 | 0.13 | 0.897   | 1.15        | -0.07378 | -0.10 | 0.919   |
| Smoking (1) | -0.44179   | -0.22 | 0.827   | 0.64        | -0.16234 | -0.12 | 0.901   |
| Stress     | 0.27745     | 1.43  | 0.152   | 1.32        | 0.18262 | 1.44  | 0.105   |

*) Significant at the \( \alpha = 5\% \)

Table 2 shows that Logit and Gompit models have \( p\text{-value} \) less than \( \alpha \). It means that age and heart rate variables are partially significant for the models. In the Logit model, determining of the influence magnitude of each significant predictor variable can be explained based on the following odds ratio.

a) Age

The increasing age of a person resulted in a person's risk of getting hypertension increased by 1.29 times.

b) Heart Rate

Each increase in heart rate resulted in a person's risk of developing hypertension increasing by 0.85 times.
4.3 Establishment Logit Model and Gompit Model

The obtained Logit and Gompit models are:

\[ g(x) = -2.004 + 0.251X_1 - 0.158X_2 - 0.931X_3 + 0.144X_4 - 0.442X_5 + 0.277X_6 \]

and

\[ g(x) = -1.329 + 0.179X_1 - 0.117X_2 - 0.885X_3 + 0.074X_4 - 0.162X_5 + 0.183X_6 \]

respectively.

When based on one of the captured data, then the value of chance factors that affect significantly to hypertension can be explained.

4.4 Test the Suitability of the Model

The suitability testing of the model is used to test the obtained Logit and Gompit models that there is no difference between the results of the observations and possible predictive results.

| Link Function | Chi-Square | p-value |
|---------------|------------|---------|
| Logit         | 3.7383     | 0.880   |
| Gompit        | 15.7938    | 0.045   |

Table 3 shows that p-value in the Logit model is greater than \( \alpha = 5\% \) compared with the p-value of the Gompit model that is less than \( \alpha = 5\% \). Therefore, it can be concluded that in this case, Logit model is more appropriate than Gompit model, i.e., there is no significant difference between observation and predicted model.

4.5 Classification Accuracy

The classification accuracy uses the following APPER (Apparent Error Rate) values.

| Link Function | APPER |
|---------------|-------|
| Logit         | 85.2% |
| Gompit        | 81.5% |

Table 4 shows that the evaluation of the Logit model is greater than the Gompit model. It means that the Logit link function is better than the Gompit link function for modelling hypertension risk factors.

5. Conclusion

The Logit link function is better than the Gompit link function for modelling hypertension risk factors. For these link functions, the significant factors that affect hypertension are heart rate and age. Based on the analysis that has been done shows that there are several obstacles that can be made as suggestions for further research that needs to be added to variables that are thought to affect the risk of hypertension. It is hoped that the community can maintain health in order to control blood pressure with a healthy lifestyle, so that it avoids hypertension and as an effort to reduce the incidence of hypertension in Indonesia.

References

[1] World Health Organization. 2011. *Hypertension Fact Sheet*. Department of Sustainable Development and Healthy Environments.

[2] Irianto, K. 2014. *Understanding Various Diseases*. Bandung: Alfabeta.

[3] Ministry of Health RI. 2017. *Center for Data and Information of Health of the Republic Indonesia of Hypertension*. INFODATIN.

[4] Ministry of Health RI. 2017. *Profile of Non-Communicable Diseases 2016*. Jakarta: Ministry of Health.

[5] Patil, M., Durairaj, M. 2005. Risk factors of hypertension among adult men: Evidence from a real world outcomes investigation in a Western Indian population. *International Journal of Advanced Research*, 3(7)pp 274-282.
[6] Khotimah. 2013. Stress as a Factor of Increasing Blood Pressure in Hypertension Patients. 
*Eduhealth Journal*, 3(2) pp 79-83.

[7] Korneliani, K., Meida, D. 2015. Obesity and Stress with Hypertension. *Journal of Public Health*, 7(2) pp117-121.

[8] Hosmer, D., Lemeshow. 2000. *Applied Logistic Regression*. USA: John Wiley and Sons.

[9] Agresti, A. 1990. *Categorical Data Analysis*. New York: John Wiley and Sons.

[10] Hardin, J.W., Hilbe, J. M. 2007. *Generalized Linear Models and Extensions*. United Stated of America: StataCorp LP.