Analysis of Hypertension Disease using Logistic and Probit Regression

D I Ruspriyanty¹ and A Sofro¹∗

¹Mathematics Department, Universitas Negeri Surabaya, Ketintang, Surabaya 60231, East Java, Indonesia
∗Corresponding Author. Email: ayuninsofro@unesa.ac.id

Abstract. Hypertension (high blood pressure) is when blood pressure has reached or exceeded 140 mmHg (systole) and 90 mmHg (diastole). Suspected factors of the disease are age, obesity, education, partner status, and occupation. In this paper, we analyze models for determining high-risk factors of hypertensive disease. The results show that education and partner status are a significant factor affecting the disease. The Akaike Information Criterion value (AIC) of the logistic model is 118.64 and the probit model is 118.79. Therefore, the performance of a logistic model is better than probit due to providing the smaller value of AIC.

1. Introduction
Hypertension is a dark killer (the silent killer) because it includes a deadly disease, because it is accompanied by symptoms that are often regarded as an ordinary disorder so that the sufferer is late to realize the coming of hypertension. The sign of hypertension are headaches, palpitations, difficulty breathing after working hard or lifting a heavy burden, fatigue, blurred vision, ringing ears (tinnitus), and nosebleeds [12]. Death due to hypertension ranks higher than other causes [3].

In 2013 showed that 25.8% of Indonesia’s population suffers from hypertension. According to the Central Java Provincial Health Office, in 2013, Semarang City ranks first for hypertension cases with hypertension prevalence of 55.6% [4]. Hypertension is an increase in systolic blood pressure of more than 140 mmHg and diastolic blood pressure of more than 90 mmHg in two measurements with an interval of five minutes in a rest/quiet state [9]. Factors that influence the occurrence of hypertension is divided into 2 groups, namely uncontrollable factors and factors that can be controlled. Uncontrollable factors such as age, sex, genetics, race and controllable factors such as exercise habits, diet, alcohol consumption, salt, coffee, and stress.

Based on the case above, we describe the hypertension disease as a dependent variable and there are five independent variables used. Thus, we use logistic and probit regression method. Logistic regression is one of the non-linear regression methods that can be used to find the response variable relationship of dichotomous (nominal or ordinal in two categories) or polychotomous (with nominal or ordinal scale with more than two categories) with one or more predictor variables [1]. The probit regression is also a nonlinear regression used to analyze the relationship between one variable response (dependent variable) with some predictor variable (independent variable), where the response variable is a qualitative data dichotomy that is 1 to state the existence of a characteristic and 0 to express the non-existence a characteristic [6].

In this research, we will analyze the factors that influence hypertension disease with logistic and probit regression. The data is collected from Budi Artiyaningrum in 2015 [4]. The status of disease from the observations become the dependent variable, i.e if the subject is hypertension, the case is categorized one, otherwise zero. In this paper, we will discuss the method, i.e logistics and probit regression. How to estimate parameter and analyze the disease data using both of those approaches.

2. Method of Research
2.1 Logistic Regression
Logistic regression is one of the non-linear regression methods that can be used to find the relationship of dichotomous (nominal or ordinal in two categories) or polychotomous (with nominal or ordinal scale...
with more than two categories) with one or more predictor variables [1]. If the response variable consists of two categories (dichotomy/binary), i.e. \( y = 1 \) (success) and \( y = 0 \) (fail), logistic regression method can be applied is binary logistic regression [2], then the response variable \( y \) follows the Binomial distribution with probability function:

\[
f(y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i},
\]

where \( y_i = 0; 1 \), \( \pi(x_i) \) is the probability of success, \( 1 - \pi(x_i) \) is the probability of failure, and \( i = 1, 2, ..., n \) is the observation index. Logistic regression models used are [5]:

\[
\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p}}, 
\]

where \( p \) is the number of independent variables. The logistic regression model in equation (2) is transformed using the logit transformation obtained:

\[
g(x) = \log \left( \frac{\pi(x)}{1-\pi(x)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.
\]

Then, it will explain the estimation of the \( \beta \) parameter in logistic regression by using Maximum Likelihood Estimation (MLE).

### 2.2 Parameter Estimation

Maximum Likelihood Estimator (MLE) is one of the methods used to estimate/predict parameters on logistic regression. The method estimates the \( \beta \) parameter by maximizing the likelihood function provided that the data must follow a certain distribution. If the dependent variable \( y \) follows the Binomial distribution with the probability of success \( \pi(x_i) \) and the probability of failure \( 1 - \pi(x_i) \) thus the likelihood function for the logistic regression model is:

\[
L(\beta) = \prod_{i=1}^{n} f(y_i; \beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}. 
\]

Mathematically it would be easier if equation (4) is brought to the log-likelihood function and maximizes the log-likelihood. The following is the log-likelihood function:

\[
\ell(\beta) = \sum_{j=0}^{p} \left[ \sum_{i=1}^{n} y_i x_{ij} \right] \beta_j - \sum_{i=1}^{n} \ln \left[ 1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}} \right].
\]

Equation (5) is differentiated to \( \beta \) to obtain the following equation:

\[
\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \pi(x_i) = 0 \quad , j = 0, 1, 2, ..., p, (6)
\]

where \( p \) is the number of predictor variables with

\[
\pi(x_i) = \frac{e^{\sum_{j=0}^{p} \beta_j x_{ij}}}{1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}}}, \quad j = 0, 1, 2, ..., p .
\]

To find the derivatives of equation (6) which are equated with zero, it often does not get explicit results so that the using the numerical method is the Newton Raphson method to optimize equation (6).

### 2.3 Probit Regression

The probit model is a non-linear model used to analyze the relationship of response/dependent variable with several predictor/independent variables where the response variable is qualitative data dichotomy that is 0 and 1 [6]. The probit regression method uses the cumulative distribution function Normal (Normal cumulative distribution function) to explain the function of the equation. Since the response/dependent variable is dichotomous/binary, the response/dependent variable \( y \) follows the Binomial distribution with the probability function as follows:

\[
f(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i},
\]
with $y_i = 0; 1$, $\pi_i$ is the probability occurrence of the i for $y_i = 1$, and $1 - \pi_i$ is the probability occurrence of the i for $y_i = 0$. The transformation function in the probit regression model is the cumulative function of the Normal distribution as a link function in GLM.

$$P(y_i = 1|x_i) = \Phi(x_i|\beta) = \int_{-\infty}^{x_i} \phi(z)\,dz, \quad (7)$$

where $\Phi(.)$ is the cumulative function of the Normal distribution and $\phi(.)$ is the Normal distribution probability function.

$$\Phi(g(x)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{g(x)} e^{-\frac{z^2}{2}} dz.$$  

In general, the probit regression model can be expressed as follows:

$$\pi_i = \Phi(Z_i) = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i).$$

Since probit model is related to the cumulative function of Normal distribution, it can be written probit model as follows:

$$Z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i.$$  

To obtain an expectation of the probit value ($Z_i$), then the inverse of the normal cumulative distribution function can be obtained:

$$Z_i = \Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i.$$  

Then, it will explain the estimation of the $\beta$ parameter in probit regression by using Maximum Likelihood Estimation (MLE).

2.4 Parameter Estimation

Maximum Likelihood Estimator (MLE) is one of the methods used to estimate/predict parameters of a known distribution model. The likelihood function for the probit regression model can be written as follows:

$$L(\beta) = \prod_{i=1}^{n} f(y_i; \pi_i). \quad (8)$$

The transformation function in the probit regression model is the cumulative function of the Normal distribution so that the equation (8) becomes:

$$L(\beta) = \prod_{i=1}^{n} (\Phi(x_i|\beta)^{y_i} (1 - \Phi(x_i|\beta))^{1-y_i},$$

where $\Phi(.)$ is the cumulative function of the Normal distribution. The log likelihood obtained as follows:

$$\ell(\beta) = \sum_{i=1}^{n} y_i \log(\Phi(x_i|\beta)) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \Phi(x_i|\beta)). \quad (9)$$

The parameter value of $\beta$ can be obtained by maximized the log-likelihood function of finding the derivative of $\beta$. To find the derivatives of equation (9) which are equated with zero, it often does not get explicit results so that the using the numerical method is the Newton Raphson method to optimize equation (9).

2.5 Testing of Parameter Significance

The model obtained needs to be tested whether the independent variables contained in the model have a significant relationship with the dependent variable. The test is a simultaneous test and a partial test [5]. The simultaneous test is performed to find out the significance of $\beta$ coefficient to the overall response/dependent variable, with the following test hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1 : \text{there is at least one } \beta_j \neq 0 \text{ with } j = 1,2,\ldots,p$$
where \( n_1 = \sum_{i=1}^{n} y_i \), \( n_0 = \sum_{i=1}^{n} (1 - y_i) \), and \( n = n_1 + n_0 \). The \( G \) test statistic follows the Chi-square distribution so that reject \( H_0 \) if \( G > \chi^2(\alpha, p) \) with \( p \) is the number of independent variables or the number of parameters in the model at a significant level \( \alpha \) or \( p\)-value < \( \alpha \) [5]. The partial test is to determine the effect of \( \beta \) coefficient individually by comparing the standard error. The hypothesis is as follows:

\[
H_0 : \beta_j = 0 \\
H_1 : \beta_j \neq 0 \text{ with } j = 1, 2, ..., p
\]

Wald test statistics [10]:

\[
W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} (11)
\]

\( W \) test statistic follows the standard normal distribution with the rejection region: reject \( H_0 \) if value \( |W| > Z_{(1-\alpha)/2} \) or \( p\)-value < \( \alpha \). \( \hat{\beta}_j \) is the value of the regression parameter estimation and \( SE(\hat{\beta}_j) \) is the standard error [2].

2.6 Best Model Selection Criteria

AIC is the model selection criteria considering the number of parameters and the smaller the AIC value the better the model. According to [4] Akaike Information Criterion (AIC) is defined as follows:

\[
AIC = -2 \ln L(\hat{\beta}) + 2k (12)
\]

where \( L(\hat{\beta}) \) is the likelihood value, \( k \) is the number of parameters.

3. Experiment Set Up

3.1 Dataset

The data used is obtained from the results of previous research with the title of factors related to the incidence of uncontrolled hypertension in patients who perform routine checks at Kedungmundu Public Health Center [4].

3.2 Operational Definition of Research Variable

The dependent variable is hypertension disease. The dependent variable is categorical, there are only two categories of response 1 states a person suffering from hypertension and response 0 states someone who does not suffer from hypertension. Independent variables are: age is the length of respondent’s life calculated from birth to the last birthday and it is a continuous variable expressed in units of years, obesity is overweight as a result of excessive body fat accumulation. How to know the level of obesity is by using Body Mass Index (BMI) [8].

\[
BMI = \frac{\text{Weight (kg)}}{\text{Height (m)}^2}
\]

education is the level of school that has been followed by the respondents which are divided into two categories are 1: low (no school, elementary, junior high), 0: high (high school, college), partner status is the condition of respondents based on the presence and absence of a life partner (husband/wife) in everyday life which is divided into two categories are 1: no couple (single, widower, widow, separate), 0: there is a couple (married), and occupation is the condition of respondents who have jobs or do not have a job in everyday life which is divided into two categories are 1: not work (housewife), 0: work (self-employed, laborers, farmers, employees).
4. Result and Discussion

4.1 Factors Affecting of Hypertension Disease Using Logistic Regression

The results simultaneously with the value of test statistic G is 15.35 which is more than critical value from Chi-square with alpha 0.05 and degree of freedom 5. We also get p-value is 0.008968 which is less than alpha five percent. Thus there is a strong evidence to reject $H_0$. So it can be concluded that there are at least one of the independent variables significant on hypertension disease. Meanwhile, the result of parameters on the partial test performed using software R is shown in table 1.

| Independent Variable | Coef.  | SE      | $z_{value}$ | $p-value$ | Exp(Coef.) |
|----------------------|--------|---------|-------------|-----------|------------|
| Age                  | -0.01226 | 0.03335 | -0.368      | 0.7132    | 0.989      |
| Obesity              | 0.06350 | 0.06210 | 1.023       | 0.3065    | 1.066      |
| Education            | 1.74148 | 0.74705 | 2.331       | 0.0197 *  | 5.706      |
| Partner Status       | 1.29371 | 0.58716 | 2.203       | 0.0276 *  | 3.646      |
| Occupation           | -0.42108 | 0.49787 | -0.846      | 0.3977    | 0.656      |
| (intercept)          | -2.62926 | 2.30484 | -1.141      | 0.2540    | 0.072      |

*) significant at $\alpha = 5\%$

Based on the result of partial test logistic regression parameter in table 1 above shows that the value of test statistic $W$ for education is 2.331 and the value of test statistic $W$ for partner status is 2.203 which are more than critical value from $Z(1-\alpha)/2$ with alpha 0.05. We also get the p-value of education is 0.0197 and p-value of partner status is 0.0276 which are less than alpha five percent. Thus the independent variables are education ($x_3$) and partner status ($x_4$) significant to the model. So it can be concluded that education and partner status are the factors that influence hypertensive disease because it has the $p-value < \alpha$ with significance level $\alpha = 5\%$. Odds ratio performed to determine the magnitude of the influence of each independent variable is significant as follows: respondents with no school and primary/junior high school have a 5.706 times greater risk of developing hypertension compared to senior high school/college education and respondents with status single/widower/widow/separate has a risk 3.646 times greater suffering from hypertension compared with a married status.

Logistic regression model is following:

$$\pi(x) = \frac{e^{-2.62926 - 0.01226x_1 + 0.06350x_2 + 1.74148x_3 + 1.29371x_4 - 0.42108x_5}}{1 + e^{-2.62926 - 0.01226x_1 + 0.06350x_2 + 1.74148x_3 + 1.29371x_4 - 0.42108x_5}}$$

4.2 Factors Affecting of Hypertension Disease Using Probit Regression

The results simultaneously with the value of test statistic G is 15.20 which is more than critical value from Chi-square with alpha 0.05 and degree of freedom 5. We also get p-value is 0.009513 which is less than alpha five percent. Thus there is a strong evidence to reject $H_0$. So it can be concluded that there are at least one of the independent variables significant on hypertension disease. Meanwhile, the result of parameters on the partial test performed using software R is shown in table 2.

| Independent Variable | Coef.  | SE      | $z_{value}$ | $p-value$ |
|----------------------|--------|---------|-------------|-----------|
| Age                  | -0.007738 | 0.020460 | -0.378      | 0.7053    |
| Obesity              | 0.037465 | 0.037046 | 1.011       | 0.3119    |
| Education            | 1.034035 | 0.434989 | 2.377       | 0.0174 *  |
| Partner Status       | 0.782626 | 0.352979 | 2.217       | 0.0266 *  |
| Occupation           | -0.236106 | 0.299571 | -0.788      | 0.4306    |
| (intercept)          | -1.540951 | 1.383035 | -1.114      | 0.2652    |

*) significant at $\alpha = 5\%$
The logistic regression model obtained is as follows:
\[
\pi_i = \Phi(Z_i) = -1.540951 - 0.007738x_1 + 0.037465x_2 + 1.034035x_3 + 0.782626x_4 - 0.236106x_5.
\]

Based on the result of partial test probit regression parameter in Table 2 above shows that the value of test statistic \( W \) for education is 2.377 and the value of test statistic \( W \) for partner status is 2.217 which are more than critical value from \( Z_{(1-\alpha)/2} \) with alpha 0.05. We also get the p-value of education is 0.0174 and p-value of partner status is 0.0266 which are less than alpha five percent. Thus the independent variables are education \( (x_3) \) and partner status \( (x_4) \) significant to the model. Based on the model of probit regression above shows that the significant factors of the disease are: respondents with low education (no school and primary/junior high) have a higher risk of hypertension with a Z score of 1.034035 compared with higher education (high school/college graduation) and respondents who have no partner status (single/widower/widowed/separate) have a higher risk of hypertension with Z score of 0.782626 compared with the status of a couple (married).

4.3 Best Model Selection Criteria

The results of AIC value is shown in Table 3.

| Model              | AIC  |
|--------------------|------|
| Logistic Regression| 118.64 |
| Probit Regression  | 118.79 |

The smaller AIC value, the better model. The AIC value of the logistic model is smaller than the one. Thus, from Table 3, it shows that logistic regression model is better than probit regression model.

5. Conclusion

Based on the results of analysis and discussion in this study can be concluded that hypertension disease is influenced by factors are education and partner status at a significance level of 5%. The logistic regression model is better than the probit regression model because the AIC value of logistic regression is the smaller than another model.

References

[1] Alan A 1990 Categorical Data Analysis (Canada: John Wiley & Sons Inc)
[2] Amr I A 2010 Applying Logistic Regression Model to the Second Primary Cancer Data (Egypt: Ain Shams University Press) pp 1-20
[3] Bambang H 2011 Hipertensi: The Silent Killer (Jakarta: Perhimpunan Hipertensi Indonesia)
[4] Budi A 2015 Faktor-Faktor yang Berhubungan dengan Kejadian Hipertensi Tidak Terkendali pada Penderita yang Melakukan Pemeriksaan Rutin di Puskesmas Kedungmendu Kota Semarang Tahun 2014 (Semarang: Universitas Negeri Semarang)
[5] David W H and Stanley L 2000 Applied Logistic Regression (New Jersey: John Wiley & Sons, Inc)
[6] Evy W 2013 Model Regresi Probit untuk Mengetahui Faktor-Faktor yang Mempengaruhi Jumlah Penderita Diare di Jawa Timur (Surabaya: Universitas Negeri Surabaya)
[7] Hamparsum B 2000 Akaike's Information Criterion and Recent developments in informational complexity Journal of Mathematical Psychology 44(1) pp. 62–91
[8] Kemenkes RI 2012 Cara Mencegah dan Mengatasi Obesitas (Jakarta: Direktorat Bina Gizi Masyarakat)
[9] Kemenkes RI 2014 Pusdatin Hipertensi. Infodatin pp. 1-7
[10] Saroje K S and Habshah M 2010 Importance of Assessing the Model Adequacy of Binary Logistic Regression Journal of Applied Sciences 48(6) pp. 479-486
[11] Sumarni H A 2016 Emisi Transportasi (Makassar: Penebar Plus+)
[12] Vitahealth 2006 Hipertensi (Jakarta: PT. Gramedia Pustaka Utama)