Using Logistic Regression to Estimate the Influence of Adolescent Sexual Behavior Factors on Students of Senior High School 1 Sangatta, East Kutai-East Kalimantan

Darnah and Memi Norhayati
Department of Mathematics, Mulawarman University, Samarinda, Indonesia, 75123
e-mail: darnah.98@gmail.com

Abstract—Adolescence is a transitional stage of physical and psychological development that generally occurs during the period from puberty to legal adulthood. Puberty is the time in which a child's sexual and physical characteristics mature. Adolescents have been found at high risk for many negative health consequences related to sexual risk-taking behavior, including infection with human immunodeficiency virus, other sexually transmitted disease, and unintended pregnancy. This research aims to determine the relationship between the reproductive health knowledge and using gadget with adolescent sexual behavior using chi- square test. Additionally, binary logistic regression was used to modeling it. Subjects in this research are students of Senior High School 1 Sangatta, East Kutai-East Kalimantan. The research shows that reproductive health knowledge has a relationship with adolescent sexual behavior. The students have reproductive health knowledge are good to have adolescent sexual behavior unrisk 2.6 more than students have reproductive health knowledge are bad.

Keywords—Adolescents, Gadget, Logistic Regression, Reproductive Health, Sexual Behavior.

INTRODUCTION

According to the regulation of the Ministry of Health Republic of Indonesia (Permenkes RI) number 25 in 2014, adolescents are resident in the age range 10-18 years and according to the Indonesian population and family information network (BKKBN), the adolescent's age range was 10-24 years old and unmarried. The World Health Organization (WHO) was defined adolescents are resident in the age range 12-24 years [1].

Adolescence is a transitional stage of physical and psychological development that generally occurs during the period from puberty to legal adulthood (age of majority). Adolescence is usually associated with the teenage years, but its physical, psychological or cultural expressions may begin earlier and later. Puberty has various definitions: generally, it is the signal of the onset of the female reproductive cycle that means the onset of puberty [2]. It is also known as a period of increased secretion of gonadal steroid hormones, therefore it is a sensitive period during the development of girl's life [3]. Another definition of puberty is the time in which a child's sexual and physical characteristics mature [4].

In recent years, professional and public attention has been directed to the numerous health risks of unsafe sexual behavior. Adolescents, have been found at high risk for many negative health consequences related to sexual risk-taking behavior, including infection with human immunodeficiency virus (HIV), other sexually transmitted disease (*e.g.*, syphilis, chlamydia), and unintended pregnancy [5]. Adolescent sexual behavior is influenced by many factors, including information technologies such as the internet, television, individual perceptions, personality characteristics, educational aspiration, *etc*.

A gadget is a tool that current technology is growing rapidly with a variety of applications that have special functions that are often abused by the adolescent. In recent years, we have seen a profound increase in the use of gadget (handphone, iPad, smartphone, tablet) by the adolescent. Adolescents, in particular, have been found to be at high risk for much negative consequence related to sexual risk-taking behavior, mental illness, attention deficit, psychological and behavior disorders. A student in the elementary school has sexual harassment against his friends because he was watched porn video [6].

According to the result of the Demography and Health Survey of Indonesia (SDKI) in 2012, 16.9 % of all adolescents in 2007 increase to 21.6 % in 2012 had sexual intercourse [7]. The Indonesian Planned Parenthood (PKBI) of East Kalimantan was cooperated with women empowerment child protection and family planning of East Kalimantan in 2009 shows that of the total 400 respondents consisting of 192 adolescent male and 208 female were 14 % had sexual intercourse when their age 10-20 years [8]. In 2013, HIV/AIDS disease and infection disease in East Kutai are the rank third of all others district in East Kalimantan, 36 people HIV/AIDS disease and 273 person infection disease [9].

In this study, we aim to determine the relationship between the reproductive health knowledge of adolescent and gadget with adolescent sexual behavior on students of senior high school 1 Sangatta, East Kutai-East Kalimantan using chi-square test. Additionally, binary logistic regression was used to estimate the influence of adolescent sexual behavior factors. Logistic regression is one of the varieties of popular multivariate tools used in biomedical informatics especially for correlating the dichotomous outcomes with the predictor variables that include different physiological data [10]. Al-Ghamdi [11] developed a logistic model and used it to estimate the influence of accident. Yusuff, et al., [10] was using logistic regression to analysis breast cancer.

Logistic regression is used to describe data and to explain the relationship between a response variable and one or more predictor variables. The response variable is discrete (binary or multinomial) and predictor variables are discrete or metric. Multinomial logistic regression analysis is capable of showing the best way to find the conclusion and be made as a parsimonious model to describe the relationship between response and predictor variables. Binary logistic regression is one of the logistic regression analysis methods whereby the predictor variables are dummy variables. Independent variables consist of different size levels whereas dependent variables must be linear and fulfills the response that is needed for this method. A logistic regression model is the result of the non-linear transformation of the linear regression model. The difference between logistic regression and linear regression is that the outcome variable in logistic regression is dichotomous [12].

In logistic regression, the predicted odd ratio of the positive outcome is expressed as a sum of the product. The product is formed by multiplying the values of predictor variable and its coefficients. The probability of the positive outcome is obtained from the odd ratio through a simple transformation [13].

Chi-Square Test

When the data of research consist of frequencies in discrete categories, the chi-square test may be used to determine the significance of differences between two independent groups. The measurement involved may be as weak as nominal scaling. The hypothesis under test is usually that the two groups differ with respect to some characteristic and therefore with respect to the relative frequency with which group members fall into several categories. To test this hypothesis, we count the number of cases from each group which falls in the various categories, and compare the proportion of cases from one group in the various categories with the proportion of

ISBN: 978-602-60569-5-5

cases from the other group. The chi-square test for two independent two sample is [14]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})\hat{}2}{E_{ij}} \quad (1)$$

Where O_{ij} represent an observed number of cases categorized in the ith row of jth column and E_{ij} is a number of cases expected under H_0 to be categorized in the ith row of jth column. The value of χ^2 yielded by equation (1) is distributed approximately as chi-square with $df = (r-1)(k-1)$, where r = the number of rows and k = the number of columns in the contingency table. To find the expected frequency for each cell (E_{ij}), multiply the two marginal totals common to a particular cell, and then divide this product by the total number of cases, N.

Logistic Regression

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more predictor variables. Quite often the outcome variable is discrete, taking on two or more possible values. The logistic regression model is the most frequently used regression model for the analysis of these data. The goal of an analysis using this model is the same as that of any other regression model used in statistics, that is, to find the best fitting and most parsimonious, clinically interpretable model to describe the relationship between a response) variable and a set of predictor variables. What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous. This difference between logistic and linear regression is reflected both in the form of the model and its assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow, more or less, the same general principles used in linear regression. Thus, the techniques used in linear regression analysis motivate our approach to logistic regression (Hosmer & Lemeshow, 2013).

In any regression analysis, the key quantity is the mean value of the response variable given the values of the predictor variable. This quantity is called the conditional mean and is expressed as:

$$E(Y|x) = \beta_0 + \beta_1 x_1$$

where Y denotes the response variable and x denotes the specific value of the predictor variable. Many distribution functions have been proposed for use in the analysis of a dichotomous response variable (Hosmer and Lemeshow, 2013).

In order to simplify notation, we use the quantity $\pi(x) = E(Y|x)$ to represent the conditional mean of Y given x when the logistic distribution is used. The specific form of the logistic regression model is:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \quad (2)$$

The transformation of the $\pi(x)$ logistic function is known as the logit transformation:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 \quad (3)$$

The importance of this transformation is that g(x) has many of the desirable properties of a linear regression model. The logit, $g(x)$, is linear in its parameters, may be continuous, and may range from -∞ to ∞, depending on the range of x. In the dichotomous variable, the value of the predictor variable given x as $y = \pi(x) + \varepsilon$. Here the quantity ε may assume one of two possible values. If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus, ε has a distribution with mean zero and variance equal to $\pi(x)[1 - \pi(x)]$. That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean, $\pi(x)$.

If Y is coded as 0 or 1 then the expression for $\pi(x)$ given in equation (2) provides (for an arbitrary value of $\beta = (\beta_0, \beta_1)$, the vector of parameters) the conditional probability that Y is equal to 1 given x. This is denoted as $\pi(x)$. It follows that the quantity $1 - \pi(x)$ gives the conditional probability that Y is equal to zero given x, $Pr(Y = 0|x)$. Thus, for those pairs (x_i, y_i), where $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$, where the quantity $\pi(x_i)$ denotes the value of $\pi(x)$ computed at x_i. A convenient way to express the contribution to the likelihood function for the pair (x_i, y_i) is the expression:

$$\pi(x_i)^{y_i}[1 - \pi(x)]^{1-y_i} \quad (4)$$

As the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in equation (4) as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\pi(x_i)^{y_i}[1 - \pi(x)]^{1-y_i} \quad (5)$$

The principle of maximum likelihood states that we use as our estimate of β the value that maximizes the expression in equation (5). However, it is easier mathematically to work with the log of equation (5). This expression, the *loglikelihood*, is defined as:

$$L(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n}[y_i \ln \pi(x_i) + (1 - y_i)n(1 - \pi(x_i))] \quad (6)$$

To find the value of β that maximizes $L(\beta)$ we differentiate $L(\beta)$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. These equations, known as the likelihood equations, are:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \quad (7)$$

and

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0 \quad (8)$$

In equations (7) and (8) it is understood that the summation is over varying from 1 to n. After the coefficients are estimated, the significance of the variables in the model is assessed. If y_i denotes the observed value and denotes the predicted value for the ith individual under the model, the statistic used in the logistic regression is:

$$SSR = SS - SSE$$

SSR = Sum of Square Regression
SS = Total Sum of Squares
SSE = Sum of Squares Error

$$SSR = [\sum_{i=1}^{n}(y_i - \bar{y}_i)^2] - [\sum_{i=1}^{n}(y_i - \hat{y}_i)^2] \quad (9)$$

Where $\bar{y}$ is the mean of the response variable. A large value suggests that the independent variable is important, whereas a small value suggests that the independent variable is not useful in explaining the variability in the response variable. The principle in logistic regression is the observed values of the response variable should be compared with the predicted values obtained from models with and without the variable in question. In logistic regression, this comparison is based on the log likelihood function defined in equation (6). Defining the saturation model as one that contains as many parameters as there are data points, the current model is the one that contains only the variable under question. The likelihood ratio is as follows:

$$D = -2\ln\left[\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}}\right] \quad (10)$$

Using equation (5) and (10), the following test statistic can be obtained:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \quad (11)$$

Where $\hat{\pi}_i = \hat{\pi}(x_i)$.

The statistic D in equation (11) is called the deviance, and it plays an essential role in some approaches to the assessment of goodness of fit. For the purpose of assessing the significance of an independent variable, the value of D should be compared with and without the independent variable in the model. The change in D due to inclusion of the independent variable in the model is obtained as follows:

$$G = D(mwov) - D(mwv)$$

mwov = for the model without the variable

mvw = for the model with the variable

The likelihood of the saturated model is common to both values of D being the difference to compute G, this likelihood ratio can be expressed as:

$$D = -2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right] \quad (12)$$

It is not appropriate here to derive the mathematical expression of the statistic G. Yet it should be said that under the null hypothesis, β_1 is equal to zero, G will follow a χ^2 distribution with one degree of freedom. Another test statistic, is the Wald statistic (W), which follows a standard normal distribution under the null hypothesis that $\beta_1 = 0$. This statistic is computed by dividing the estimated value of the parameter by its standard error:

$$W = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})} \quad (13)$$

The Wald test behaved in an aberrant manner, often failing to reject the null hypothesis when the coefficient was significant, so that the likelihood ratio test should be used in suspicious cases.

MATERIALS AND METHODS

In order to conduct a school-based questionnaire in Senior High School 1 Sangatta, East Kutai-East Kalimantan. This project acquired the support of the Faculty of Mathematics and Natural Sciences, Mulawarman University. We conducted a cross-sectional study among 100 students. These students were selected randomly using stratified random sampling. We gave each student a general overview of the study and explained the procedures. In addition, we reassured the students that their information would be protected and be strictly confidential.

The data for this research was collected the August 2016. The questionnaire included gender, age, the response variable (adolescent sexual behavior), and the predictor variables (reproductive health knowledge and using gadget by students). The description and levels of these variables are given in Table 1. The scores were computed depending on the scale, and the percentage age is calculated accordingly. Once all the data was collected, we entered the data into the Statistical Package for the Social Sciences computer software (SPSS for Widows, Version 20,0, SPSS, Inc., 2011) for analysis.

Table 1. Description of the study variables

Number	Description	Code/Value
1	Adolescent sexual behavior	0: unrisk 1: low risk
2	Reproductive health knowledge	0: good 1: bad
3	Using gadget by students	0: < 3 hours/day 1: ≥ 3 hours/day

The chi-square test was applied in order to estimate the relationship between all variables. Based on the Chi Square analysis, the P-value that is less than or equal to 0.05 is considered significant. Additionally, binary regression analysis was used to estimate the influence of adolescent sexual behavior factors.

RESULTS AND DISCUSSION

We begin this section with descriptive statistics of the adolescent sexual behavior data. The results in the sample of 100 students in Table 2 showed the percentage of gender on students of Senior High School 1 Sangata East Kutai-East Kalimantan in this research, 41 % are male and 59 % are female.

Table 2. The Percentage of gender on students of Senior High School 1 Sangata, East Kutai-East Kalimantan in Research

Variable	Category	N	%
Gender	Male	41	41
	Female	59	59
	Total	100	100

The range age of students are 16.5±0.73 years, the minimum age is 15 years and the maximum age is 19 years (Table 3).

Table 3. The mean age (years) and using gadget of students

Variable	N	Mean	SD	Min	Max
Age	100	16.5	0.73	15	19

Among 100 students in this research, 80% of them have lovers and 20% do not. But, all students had not sexual intercourse (Table 4).

Table 4. The percentage of have lovers and sexual intercourse on students

Variable	Category	N	%
Have Lovers	Yes	80	80
	No	20	20
	Total	100	100
Sexual Intercourse	Yes	0	0
	No	100	100
	Total	100	100

According to the result in Table 5, 54% of students on the senior high school 1 Sangatta East Kutai-East Kalimantan have reproductive health knowledge are good, 18% of them have adolescent sexual behavior are unrisk and 36% are low risk. Among 46% of students have reproductive health knowledge are bad, 26% of them have adolescent sexual behavior are unrisk and 20% are low risk.

Table 5. The cross tabulation between response variable and predictor varibles

Variable	Adolescent sexual behavior			
	Category	Unrisk	Low Risk	Total
Reproductive health knowledge	Good	18	36	54
	Bad	26	20	46
	Total	44	56	100
Using gadget by students	< 3 hours/day	25	23	48
	≥ 3 hours/day	19	33	52
	Total	44	56	100

Among 48 % of students using gadget below three hours every day, 25% of them have adolescent sexual behavior are unrisk and 23% are low risk. 52% of students using gadget three or more hours every day, 19% of them have adolescent sexual behavior are unrisk and 33% are low risk (Table 5).

Table 6. Signaficant values for each independent variable using Chi-Square Test

Varibles	Significat Values
Reproductive health knowledge	0.033
Using gadget by students	0.173

Table 6 shows the probability value each independent variable using chi-square test. The variables that significant values less than or equal to 0.05 are considered to be taken into the logistic regression model. The variable that had significant value 0.033 is reproductive health knowledge. Thus, this variable is considered in the logistic regression analysis.

Table 7. Paramater estimation of the Logistic Regression Model

Parameter	Estimation (B)	SE	Wald	p-value	Exp(B)
$\beta_1(1)$	0.956	0.414	5.314	0.021	2.600
Intercept	-0.262	0.297	0.778	0.378	0.769

The result of analysis adolescent sexual behavior data using logistic regression model has shown in Table 7. Parameter estimation of reproductive health knowledge variable had significant value 0.021 and odd ratio value 2.600. It's the meaning that students have reproductive health knowledge are good to have adolescent sexual behavior unrisk 2.6 more than students have reproductive health knowledge are bad.

CONCLUSIONS

80% of students have lovers and 20% do not. All students in research had not sexual intercourse.

According to the chi-square test, the variable has the relationship with adolescent sexual behavior is reproductive health knowledge. Thus, this variable considered in the logistic regression analysis. Logistic regression model shows that students have reproductive health knowledge are good to have adolescent sexual behavior unrisk 2.6 more than students have reproductive health knowledge are bad.

ACKNOWLEDGEMENTS

The authors would like to thank the Senior High School 1 Sangatta, East Kutai-East Kalimantan which participated in this research for its invaluable feedback and cooperation of their students. Also, the authors would like to thank Dr. Eng.Idris Mandang, M.Si, Dean of the Faculty of Mathematics and Natural Sciences, Mulawarman University for supporting this research.

REFERENCES

[1] Kemenkes RI, "Situasi Kesehatan Reproduksi Remaja", Jakarta: Kemenkes RI, 2013.

[2] Forman, M. R., Mangini, L, D., Thelus-Jean, R. and Hayward, M, D., "Life-course origins of the ages at menarche and menopause", Adolescent Health, Medicine and Therapeutics, pp. 41–21, 2013.

[3] Zehr, J.L., Culber, K.M., Sisk, C.L., and Klump, K.L., "An association of early puberty with disordered eating and anxiety in a population of undergraduate women and men", Hormones and Behavior, vol. 52, pp. 427-435, 2007.

[4] Ostojic, D., "Effects of Puberty Onset on Attention Deficit / Hyperactivity Disorder (ADHD) Symptoms in Female University Students", Ontario, Canada: Windsor, 2013.

[5] Kotchick, B., Shaffer, A., Forehand, R. and Miller, K., "Adolescent sexual risk behavior: A multi-system perspective", Clinical Psychology Review, vol. 21, isuue 4, pp. 493–519, 2001.

[6] Manumpil, B., Ismanto, Y. and Onibala, F., "Hubungan Penggunaan Gadget dengan Tingkat Prestasi Siswa di SMA Negeri 9 Manado", eJournal Keperawatan, vol. 3, pp. 2, 2015.

[7] BPS, "Survei Demografi dan Kesehatan Indonesia 2012", Jakarta: BPS, 2013.

[8] Taufik, A., "Persepsi Remaja Terhadap Perilaku Seks Pranikah (Studi Kasus SMK Negeri 5 Samarinda)", eJournal Sosiatri-Sosiologi, vol. 1, issue 1, pp. 31-44, 2013.

[9] Dinkes Prov. Kaltim, "Profil Kesehatan Kalimantan Timur Tahun 2013", Samarinda: Dinkes Prov. Kaltim, 2013.

[10] Yusuff, H., Mohamad, N., Ngah, U. K. and Yahaya, A.S., "Breast Cancer Analysis Using Logistic Regression", IJRRAS, vol. 10, issue 1, pp. 14-22, 2012.

[11] Al-Ghamdi, A. S., "Using logistic regression to estimate the influence of accident factors on accident severity", Analysis and Prevention (AAP), vol. 34, issue 6, pp. 729-741, 2002.

[12] Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X., "Applied Logistic Regression", Third Edition, New York: John Wiley and Sons, 2013.

[13] Samanta, B., Bird, G. L., Kuijpers, M., Zimmerman, R. A., Jarvik, G. P., Wernovsky, G., Clancy, R. R., Licht, D. J., Gaynor, J. W. and Nataraj, C., "Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms", Artificial Intelligence in Medicine, vol. 46, issue 3, pp. 201-215, 2009.

[14] Bolboaca, S. D., Jantchi, L., Sestras, A. F., Sestras, R. E. and Panfil, D.C., "Pearson- Fisher Chi-Square Statistic Revisited", Information, vol. 2, pp. 528-545, 2011.