Advances in Social Science, Education and Humanities Research (ASSEHR), volume 285
2nd International Conference on Mathematics and Mathematics Education 2018 (ICM2E 2018)

Determination of Sample Size on Logistic Regression for Sakernas Data in Jayapura Regency in 2015

1st Fadhilah Fitri
Mathematics Department, Faculty of Mathematics and Natural Sciences
Universitas Negeri Padang
Padang, Indonesia
fadhilahfitri@fmipa.unp.ac.id

2nd Bertho Tantular
Statistics Department, Faculty of Mathematics and Natural Sciences
Universitas Padjadjaran
Bandung, Indonesia
bertho@unpad.ac.id

***Abstract*—Logistic regression can be used to analyze the relationship between the factors that affect the working hours whether it more, less or normal hours. The limitation of this research is to know the population with normal working hours in Jayapura Regency in 2015. To examine the number of normal working hours and underemployment, a sampling design and sample size are needed. The variable that will be used in determining the sample size is only 1 variable, age. Furthermore, based on the information obtained from the logistic regression analysis that has been done, it will be determined the minimum sample size that going to be taken using the method proposed by Whittemore (1981), obtained N = 184 and the method proposed by Hsieh et al. (1998), obtained N = 442.**

Keywords— Logistic Regression, Sampling

I. INTRODUCTION

National Labor Force Survey (in Bahasa: "Sakernas") is the main source of employment data in Indonesia. Sakernas collects basic information related to employment that is regarding individual information from each household member aged 10 years and over. However, the information presented is only information from residents aged 15 years and over [1]. This is in accordance with the definition of working in Sakernas, that is a population aged 15 years and over who conducts activities to earn income or help earn income in the form of money or goods carried out at least 1 consecutive hour in the previous week before enumeration.

The concept and definition of labor indicators used in Sakernas refer to the concept of The Labor Force Concept (ILO) with the intention that the employment data obtained can be compared internationally [2]. The concept divides the population into 2 groups, working-age groups, and non-working age groups. The working age population is divided into the labor force and persons outside the labor force. The labor force is divided into the working population and unemployment (Haussmanns, 1992 in [2]).Then the working population will be divided into 2 categories based on the number of working hours, that are residents who work less than 35 hours per week and residents who work 35 hours or more. Residents who work less than normal hours are called underemployed [2].Residents with normal working hours and underemployed residents will be the response in this study.

A person's working hours are influenced by several factors. The number of working hours is influenced by employment status, type of work and business field [3]. Whereas according to Bukit and Bakir (1984) in [2], one of the variables that affect the number of hours worked is education. This is due to an increase in one's ability and expertise in education. High productivity because the skills possessed will increase working time. The relationship between the factors that affect the working hours whether it more, less or normal hours will be analyzed using logistic regression. The analysis was conducted to see which factors had significant statistical influence. Before the analysis is carried out, Sakernas data is first converted into binary data and then an analysis is carried out and a model for this case will be obtained.

The scope of the area in this study was Jayapura Regency. The limitation of this research is to know the population with normal working hours in Jayapura in 2015.A sampling design and sample size are needed to examine the number of normal and underemployed working hours. The variables that will be used in determining the sample size are only 1 variable. Furthermore, based on the information obtained from the logistic regression that has been conducted, it will be determined what the minimum sample size will be taken using the method [4] and [5].

II. DETERMINING SAMPLE SIZE TO ESTIMATE LOGISTIC REGRESSION PARAMETERS

A. *The Method Proposed by Whittemore (1981)*

In the calculation, Whittemore added an assumption called the probability of a small response. The purpose of this assumption is $1 + e^{\beta_0+\beta_1 X} \cong 1$. Technically, if X has a normal distribution, then the probability of a small response is $E\left(1 + e^{\beta_0+\beta_1 X}\right) = 1 + e^{\beta_0+\frac{\beta_1^2}{2}} \cong 1$. The conditions for this small response probability cause restrictions on β_0 and β_1. When X has a standard normal distribution, the formula for sample size is

$$N = \frac{(Z_\alpha + e^{-\frac{A^2}{4}} Z_\beta)^2}{e^{\beta_0} A^2}$$

To overcome the small response probability assumption, [4] proposed a modification to the equation to obtain the sample size as follows

$$N = \frac{(Z_\alpha + e^{-\frac{A^2}{4}} Z_\beta)^2}{e^{\beta_0} A^2} \times \left[1 + 2^{e^{\beta_0}} \times \frac{\left[1 + (1 + A^2) e^{\frac{5A^2}{4}} \right]}{(1 + 2e^{\beta_0})} \right]$$

B. The Method Proposed by Hsieh et al. (1998)

In this method, the idea is that the problem of logistic regression can be seen as a two-sample problem. [5] assumes that under H_1 then ${\sigma_1}^2 = {\sigma_1}^2 = \sigma^2$ and $\beta_1 = A \cong \frac{\mu_1 - \mu_2}{\sigma}$, and using a two-sample problem framework, the following sample size formula is obtained:

$$N = \frac{(Z_\alpha + Z_\beta)^2}{P^*(1 - P^*)A^2}$$

where

$$P^* = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

The advantage of this method compared to Whittemore (1981) is the assumption of small response probability is not used in calculations. The sample size when binary X covariates, in other words, has a Bernoulli distribution (π) is

$$N = \frac{\left\{ Z_\alpha \sqrt{\frac{P(1-P)}{\pi}} + Z_\beta \sqrt{\left[P_1(1-P_1) + P_2(1-P_2) \frac{1-\pi}{\pi} \right]} \right\}^2}{(P_1 - P_2)^2 (1 - \pi)}$$

where

$$P_1 = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$P_2 = \frac{e^{\beta_0 + A}}{1 + e^{\beta_0 + A}}$$

$$P = (1 - \pi) P_1 + \pi P_2$$

III. METHODS

A. Data and Variables

The data used in this research are secondary data obtained from the National Labor Force Survey (Sakernas) in Jayapura in 2015.

B. Variables

The variables that will be used to perform logistic regression analysis is the number of working hours (Y), age (X_1), education (X_2), marital status (X_3), jobs field (X_4), jobs status (X_5) andgender (X_6). The data obtained from Sakernas was first transformed into binary data with the following conditions:

Y = 0, if the working hours are less than 35 hours per week
1, if the working hours are equal to 35 hours per week or more
X_1 = 0, if the resident is 57 years old or older
1, if the resident is 15-56 years old
X_2 = 0, if the last education is elementary or lower
1, if the last education in junior high and above
X_3 = 0, if married
1, if single, divorced, widowed
X_4 = 0, if working in agriculture
1, if working not in agriculture
X_5 = 0, if the employment status is not a worker/employee
1, if the employment status is a worker/employee
X_6 = 0, if female
1, if male

Furthermore, 1 variable will be chosen to be used in determining the sample size.

IV. THE RESULT

A. The Logistic Regression Model for the Case of the Number of Working Hours of Jayapura Regency Resident in 2015

1) Model and Parameter Significance Test

Based on the results of processing using software R, the formation of logistic regression models, partial parameter tests and the fitting of a model obtained are as follows:

TABLE I. MODEL FORMATION PROCESSING RESULT AND PARAMETER SIGNIFICANCE TEST

	Estimate	**Standard Error**	**Z value**	**p-value**
(1)	***(2)***	***(3)***	***(4)***	***(5)***
(Intercept)	-2.6477	0.4588	-5.771	7.87e-09
X_1	0.6994	0.3612	1.937	0.052805
X_2	-1.1122	0.3432	-3.240	0.001194
X_3	-0.7361	0.2187	-3.366	0.000763
X_4	3.7235	0.2565	14.517	< 2e-16
X_5	1.1525	0.2625	4.390	1.14e-05
X_6	1.3788	0.2198	6.273	3.55e-10

Based on Table 1 at column (5) it can be seen that with the significant level of 95%, all variables that are age, last education, marital status, jobs field, jobs status and gender significantly influence the working hours variable of the resident (Y). Then, the logistic regression model is:

$$log\left[\frac{\pi(x)}{1 - \pi(x)}\right] = -2.6477 + 0.6994 X_1 - 1.1122 X_2 - 0.7361 X_3 + 3.7235 X_4 + 1.1525 X_5 + 1.3788 X_6$$

2) Logistic Regression Model for the Effect of Types of Jobs Field (X_4) on Working Hours

Based on the results of Praptono's research (2009) [2], the variable that has the most significant effect on the logistic regression model for Sakernas data in West Java Province in 2007 is the type of jobs field. The results obtained in this study are in accordance with that study. Hence, modeling will be conducted for variable types of jobs field (X_4). The model obtained is as follows:

TABLE II. MODEL FORMATION PROCESSING RESULT AND PARAMETER SIGNIFICANCE TEST FOR X_4

	Estimate	**Standard Error**	**Z value**	**p-value**
(1)	***(2)***	***(3)***	***(4)***	***(5)***
(Intercept)	-2.5730	0.1683	-15.29	< 2e-16
X_4	4.3005	0.2098	20.50	< 2e-16

Therefore, the model is:

$$log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -2.5730 + 4.3005\,X_4$$

3) Logistic Regression Model for the Effect of Age (X_1) on Working Hours

Furthermore, modeling was carried out for variables of age (X_1) and working hours (Y), obtained the following model:

TABLE III. MODEL FORMATION PROCESSING RESULT AND PARAMETER SIGNIFICANCE TEST FOR X_1

	Estimate	Standard Error	Z value	p-value
(1)	*(2)*	*(3)*	*(4)*	*(5)*
(Intercept)	-0.8035	0.2197	-3.658	0.000255
X_1	0.6407	0.2293	2.795	0.005193

Then, the logistic regression model is

$$log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -0.8035 + 0.6407\,X_1$$

B. Sample Size for the Case of the Number of Working Hours of Jayapura Regency Resident in 2015

Let $\alpha = 0.05$ and $\beta = 0.10$, hence $Z_\alpha = 1.96$ and $Z_\beta = 1.28$.

1) Using Whittemore Method (1981)

In Whittemore method, there is the restriction when using the formula, that is only for $0.4 \le e^A \le 2.5$ or equal to $-0.916 \le A \le 0{,}916$. If the restriction is ignored, the calculated sample size will be odd [6]. The proof is as follows

The calculated sample size when using the formula proposed by Whittemore (1981) by using the types of jobs field variable, obtained:

$$N = \frac{(1.96 + e^{-\frac{4.3005^2}{4}}1.28)^2}{e^{-2.5730}4.3005^2} \times \left[1 + 2^{e^{-2.5730}} \times \frac{\left[1 + (1 + 4.3005^2)e^{\frac{5\times 4.3005^2}{4}}\right]}{(1 + 2e^{-2.5730})}\right]$$

$$N = 2.7296 \times 1.0404e+11 = 2.8399e+11$$

The sample size by using Whittemore formulation shows an odd result, 2.8399e+11. Refer to Whittemore method, the logistic regression model that will be used is the effect of age on working hours. Therefore, the sample size is:

$$N = \frac{(1.96 + e^{-\frac{0.6407^2}{4}}1.28)^2}{e^{-0.8035}0.6407^2} \times \left[1 + 2^{e^{-0.8035}} \times \frac{\left[1 + (1 + 0.6407^2)e^{\frac{5\times 0.6407^2}{4}}\right]}{(1 + 2e^{-0.8035})}\right]$$

$$N = 183.27007$$

The calculated sample size is 183.27007, or when rounding up, the sample size is 184.

2) Using Hsieh Method (1998)

The logistic regression model that will be used is the effect of age on working hours The π value obtained is 0.4453. Then, the sample size is

$$P_1 = \frac{e^{-0.8035}}{1 + e^{-0.8035}} = 0{,}3093$$

$$P_2 = \frac{e^{-0.8035+0.6407}}{1 + e^{-0.8035+0.6407}} = 0{,}4594$$

$$P = (1 - 0.4453)P_1 + 0.4453P_2 = 0{,}3761$$

Then,

$$N = \frac{\left\{1.96\sqrt{\frac{0{,}3761\,(1-0{,}3761)}{0.4453}} + 1.28\sqrt{\left[0{,}3093(1-0{,}3093) + 0{,}4594(1-0{,}4594)\,\frac{1-0.4453}{0.4453}\right]}\right\}^2}{(0{,}3093 - 0{,}4594)^2(1-0.4453)}$$

$$N = 441{,}3142$$

It turns out that the sample size is 441.3142, or if it is rounded up, the sample size is 442.

V. CONCLUSION

Based on the study that has been done, it is concluded that:

1. The sample size in logistic regression between types of employment and the number of working hours using the Whittemore formulation (1981) turned out to show odd results, 2.8399e+11. This occurs due to restrictions on the use of the formulas, that is only for $0.4 \le e^A \le 2.5$ or equal to $-0.916 \le A \le 0{,}916$.
2. The minimum sample size in the logistic regression model between age and number of working hours using the Whittemore formulation (1981) is 184.
3. The minimum sample size in the logistic regression model between age and number of working hours using the Hsieh method (1998) is 442.

REFERENCES

[1] BPS, "Indonesia-Survey Angkatan Kerja Nasional 2015 Semester 2", Jakarta, 2016.

[2] Praptono. Agus, "Penerapan Metode Regresi Logistik pada Sampling Kompleks", Thesis, Bandung: Universitas Padjadjaran, 2009.

[3] Manning. Chris, "Kegiatan Ekonomi Angkatan Kerja di Indonesia", Pusat Penelitian Kependudukan, Yogyakarta: Universitas Gajah Mada, 1984.

[4] Whittemore. A, "The Sample Size for Logistic Regression with Small Response Probability", Journal of the American Statistical Association, Volume 76, 1981, pp. 27-32.

[5] Hsieh. F.Y., Bloch. D. A., and Larsen, M.D., "A Simple Method of Sample Size Calculation for Linear and Logistic Regression". Statistics in Medicine, Volume 17, 1998, pp. 1623-1634.

[6] Alam. M. Khorshed, M. Bhaskara Rao, dan Fu-Chih Cheng. "Sample Size Determination in Logistik Regression", The Indian Journal of Statistics, Volume 72-B, 2010, pp. 58-75.