BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1355–1364
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1355-1364

THE ORDINAL LOGISTIC REGRESSION MODEL WITH SAMPLING WEIGHTS ON DATA FROM THE NATIONAL SOCIO-ECONOMIC SURVEY

Reni Amelia[1*], Indahwati[2], Erfiani[3]

[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University
Dramaga Campus, Bogor, 16680, West Java, Indonesia

Corresponding author's e-mail: [1] reniamelia@apps.ipb.ac.id*

Abstract. *Ordinal logistic regression is a method describing the relationship between an ordered categorical response variable and one or more explanatory variables. The parameter estimation of this model uses the maximum likelihood estimation having assumption that each sample unit has an equal chance of being selected, or using a simple random sampling (SRS) design. This study uses data from the National Socio-Economic Survey (SUSENAS) having two-stage one-phase sampling (not SRS). So, the parameter estimation should consider the sampling weights. This study describes the parameter estimation of the ordinal logistic regression with sampling weight using the pseudo maximum likelihood method, especially in SUSENAS sampling design framework. The variance estimation method uses Taylor linearization. This study also provides numerical examples using ordinal logistic regression with sampling weight. Data used is 121,961 elderly spread over 514 districts/cities. Testing data (20%) is used to obtain the accuracy of the prediction results. The variables used in this study are the health status of the elderly as the response variable and nine explanatory variables. The results of this study indicate that the ordinal logistic regression model with sampling weights is more representative of the population and more capable to predict minority categories of the response variable (poor and moderate health status) than is without sampling weights.*

Keywords: *ordinal logistic regression, pseudo maximum likelihood, sampling weight, SUSENAS, Taylor linearization.*

Article info:

Submitted: 22th July 2022 *Accepted: 20th October 2022*

How to cite this article:
R. Amelia, Indahwati, and Erfiani, "SAMPLING WEIGHTS IN THE ORDINAL LOGISTIC REGRESSION MODEL ON DATA FROM THE NATIONAL SOCIO-ECONOMIC SURVEY", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1355-1364, Dec., 2022.

1. INTRODUCTION

Ordinal logistic regression is a method that describes the relationship between ordered categorical response variables that have more than two categories and one or more explanatory variables [1]. Estimating the parameters of the ordinal logistic regression model use the Maximum Likelihood Estimator (MLE) method. The assumption of the model parameter estimation using MLE is that each sample unit has an equal chance of being selected [2]. In fulfilling the random assumption, the sampling method must use a simple random sampling (SRS) technique.

The National Socio-Economic Survey (SUSENAS) is a survey conducted by BPS- Statistics Indonesia twice in one year (March and September). Currently, SUSENAS is the main support for meeting the needs of the government in implementing national development in line with The National Medium-Term Development Plan for 2020-2024 and international development goals (Sustainable Development Goals/SDGs) [3]. This research focuses on the use of sampling weight for the March 2020 SUSENAS. The sampling design of this survey did not have SRS, but two stages one phase sampling. Sampling was done by taking samples of the census blocks in the first stage and taking samples of households in the selected census blocks in the second stage. This SUSENAS sampling design makes each sample unit have an unequal chance of being selected so calculating the parameter estimator needs to be made using sampling weights.

[4] used binary logistic regression with sampling weights for the classification of the elderly based on their level of difficulty in daily activities. The total sample used was 12.769 elders from the 2001 Medicare Current Beneficial Survey (MCBS) data in the United States. The sampling method on MCBS was multistage sampling. The results show that the use of sampling weights is more representative of the population than without sampling weights. The relationship between independent variables and explanatory variables is also able to reflect the actual population.

The use of sampling weights in logistic regression was also carried out by [5]. Their research aims to predict the factors that influence the use of mosquito nets in Mozambique. The data used were sourced from the Mozambique Demographic Health and Survey (MDHS) 2011. The MDHS used a multistage sampling design. The total sample used was 13,745 women of productive age (15-45 years). The results showed that the use of sampling weights produces reliable results. The resulting estimator was more efficient because it takes into account the effects of complex sampling designs.

[6] used pilot weights in binary and multinomial logistic regression models. His research also compared the results of classifying a person's body mass index using quasi-likelihood and correct likelihood methods (considering sampling weights). The total sample used was 265 Nairobi Hospital patients. The results showed that there were some similarities and differences between the quasi-likelihood method in parameter estimation, standard error, and p-value statistics. The correct likelihood method was believed to be able to explain better data where each observation did not have an equal chance of being selected.

The advantage of this study compared to previous studies is that this study can provide an in-depth understanding of the use of sampling weights in the ordinal logistic regression model, especially for data which each observation does not have an equal chance of being selected. The data used in this study is from SUSENAS March 2020. This SUSENAS data is widely used by the government and various groups for policy-making and research purposes. To make it easier to understand, in this paper, ordinal logistic regression with sampling weight will be directly applied to the March 2020 SUSENAS data with the observation unit is the elderly in Indonesia and the dependent variable is the health status of the elderly. The results of this study are expected to be an alternative method used when analyzing data which response variables are ordinal by considering sampling weight.

2. RESEARCH METHODS

This research is an exploratory study by reviewing several journals, books, conference papers, and other sources on the internet and BPS-Statistics Indonesia. The model used in this paper is ordinal logistic regression or commonly known as the cumulative logit model [1] with sampling weight. The sampling weight used the March 2020 SUSENAS for each individual data. To make it easier to understand, this paper uses a numerical example. The data used is 121.961 elderly spread across 514 regencies/districts/cities in

Indonesia. This data source is from the March 2020 SUSENAS held by BPS-Statistics Indonesia.

3. RESULTS AND DISCUSSION

3.1. SUSENAS Sampling Weight

The March 2020 SUSENAS was focused on district/regency/city level estimation so that the sampling design used was applied to each district/regency/city separately [3]. The SUSENAS March 2020 sampling design used two stages one phase sampling. Table 1 explains that in the first stage the census block was selected. From about 720 thousand census blocks, 40 percent of the census blocks were taken using a probability proportional to size (PPS) method using the size of the number of households resulting from the 2010 Population Census (SP2010) in each stratum in the district/regency/city. After that, a number of n census blocks were selected according to systematic sampling in each urban/rural stratum in each district/regency/city. The next stage was to select 10 households that have been updated systematically.

Table 1. The Sampling Method of March 2020 SUSENAS

Phase	Unit	Number of Unit in Stratum h		The Sampling Method	Sampling Probability	Sampling Fraction
		Population	Sample			
1	The census block (1)	N_{kh}	n'_{kh} (40% census block)	*PPS-with replacement*	$\frac{M_{khi}}{M_{kh}}$	$n'_{kh}\frac{M_{khi}}{M_{kh}}$
	The census block (2)	n'_{kh}	n_{kh} (according to allocation)	Systematic	$\frac{1}{n'_{kh}}$	$\frac{n_{kh}}{n'_{kh}}$
2	The households	$M^{up}{}_{khi}$	$\bar{m}$ (10 households)	Systematic	$\frac{1}{M^{up}{}_{khi}}$	$\frac{\bar{m}}{M^{up}{}_{khi}}$

***Data source**: BPS-Statistics Indonesia*

Equation (1) is the SUSENAS sampling fraction.

$$f = f_1\, x\, f_2\, x\, f_3 = n'_{kh}\frac{M_{khi}}{M_{kh}}\, x\, \frac{n_{kh}}{n'_{kh}}\, x\, \frac{\bar{m}}{M^{up}{}_{khi}} = \frac{n_{kh}M_{khi}\bar{m}}{M_{kh}M^{up}{}_{khi}} \quad (1)$$

where N_{kh} is the number of census block in stratum h and district/regency/city k, n'_{kh} is 40% of the number of census blocks in stratum h district/regency/city k, n_{kh} is the number of census blocks sample in stratum h district/regency/city k, M_{khi} is the number of household in the census blocks i stratum h district/regency/city k (2010 Population Census data), M_{kh} is the number of household in the stratum h district/regency/city k (2010 Population Census data), $M^{up}{}_{khi}$ is the number of household in the census blocks i stratum h district/regency/city k (updating result), $\bar{m}$ is the number of households sample in each census block, k is district/regency/city ($k = 1, 2, \cdots, K$), h is stratum ($h = 1, 2, \cdots, H_k$), i is census block ($i = 1,2, \cdots, n_{kh}$), m is household ($m = 1, 2, \cdots, \bar{m}$), f_1 is sampling fraction of census block sampling (first stage), f_2 is sampling fraction of census block sampling (second stage), and f_3 is sampling fraction of household sampling.

From the sampling design, an initial/base weight is obtained which describes the sampling opportunity. The basic weight of SUSENAS is the inverse of the sample fraction. Equation (2) is the formula of SUSENAS basic weight.

$$W^{design} = \frac{1}{f} \quad (2)$$

where f is sampling fraction. During the implementation in the field, it was difficult for the enumerator to obtain all the desired information. To compensate for non-response and non-coverage, adjustments were made to the basic weights. The final weight was denoted by W^{final}. W^{final} is the sampling weight for the household.

Calculation of the SUSENAS sampling weight by BPS-Statistics Indonesia was carried out up to the household level, not to the household members level (individuals data). The unit of observation in this study is a individual, not the household. The sampling weight for individuals data is adjusted to the sampling weight for each household. If there are more than one individual in the same household, they have the same sampling weight. The sampling weight of each individual denoted by w_{khimo}, which means the sampling weight of the observation unit o household m census block i stratum h and district/regency/city k, o is observation unit (individual) $(o = 1, 2, \dots, O_{khim})$.

3.2 The Ordinal Logistic Regression Model with Sampling Weights

[7] formulate an ordinal logistic regression model with sampling weight especially based on stratified cluster sampling design. The ordinal logistic regression model with sampling weight in this study is based on [7] formula with some modifications in line with SUSENAS sampling weight. In the SUSENAS, the sampling method was two stage one phase sampling [3]. Denote the cumulative sum of the expected proportions for the first d categories of variable Y by $F_{khimod} = \sum_{t=1}^{d} \pi_{khimot}$ for $d = 1,2,\cdots,D$. Then the link function used in the cumulative logit which is stated in Equation (3).

$$\log\left(\frac{F_{khimod}}{1 - F_{khimod}}\right) = \alpha_d + \boldsymbol{x}_{khimo}\boldsymbol{\beta} \tag{3}$$

where $\boldsymbol{\pi}_{khnmo} = E(\boldsymbol{y}_{khimo}|\boldsymbol{x}_{khimo}) = (\pi_{khimo1}, \pi_{khimo2}, \dots, \pi_{khimoD})'$, $\boldsymbol{\pi}_{khimo}$ is vector of the expected value of the response Y, $\boldsymbol{y}_{khnmo}$ is a line vector with dimension D, its elements are the value of the Y (response variable) in the first D category, $\boldsymbol{x}_{khimo}$ is k-dimensional row vector for the explanatory variable of the observation unit o household m census block i stratum h and district/regency/city k. If there is an intercept, then x_{khim1} is 1. $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)'$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_D)', \alpha_1 < \alpha_2 < \cdots < \alpha_D, \boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')', \boldsymbol{\theta}$ is p-dimensional column vector for the regression coefficient.

Determine the following D-dimensional column vector:

$$\boldsymbol{q}_{khimo} = ((F_{khimo1}(1 - F_{khimo1})), (F_{khimo2}(1 - F_{khimo2})), \dots, (F_{khimoD}(1 - F_{khimoD})))'$$

For example, U is a matrix $D\ X\ D$. $U = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \dots & \dots & \\ & & & 1 & -1 \\ & & & & 1 \end{bmatrix}$

The first partial derivative of the matrix is $\boldsymbol{D}_{khimo} = \begin{bmatrix} diag(\boldsymbol{q}_{khimod})U \\ \boldsymbol{q}'_{khimo}U \otimes \boldsymbol{x}'_{khimo} \end{bmatrix}$, $\otimes$ is *Kronecker Product.*

The evaluation of the Pseudo Maximum Likelihood Estimator $\widehat{\boldsymbol{\theta}}$ is stated in Equation (4):

$$\boldsymbol{e}_{khim.} = \sum_{o=1}^{O_{khim}} w_{khimo} \begin{pmatrix} diag(\widehat{\boldsymbol{q}}_{khimo})\boldsymbol{U}(diag(\widehat{\boldsymbol{\pi}}_{khimo}))^{-1} \\ \boldsymbol{\tau}_{khimo} \otimes \boldsymbol{x}'_{khimo} \end{pmatrix} (\boldsymbol{y}_{khimo} - \widehat{\boldsymbol{\pi}}_{khimo}) \tag{4}$$

where

$\boldsymbol{\tau}_{khimo} = (\widehat{\boldsymbol{q}}'_{khimo})\boldsymbol{U}(diag(\widehat{\boldsymbol{\pi}}_{khimo})^{-1} + \hat{\pi}_{D+1}^{-1}\hat{q}_D\boldsymbol{1}', \boldsymbol{1}$ is a D-dimensional column vector whose elements are 1.

In [8], it is said that the important thing to do in a logistic regression model is to test the significance of the model parameters. There are two parameter tests used, namely the G test (Likelihood Ratio Test) and the Wald test. The G test is used to determine the effect of all explanatory variables in the model together. Wald test is used to test the significance of each parameter.

3.3 Pseudo Maximum Likelihood

The Pseudo Maximum Likelihood (PML) estimation method is a development of the maximum likelihood method [9]. The PML had a smaller bias than maximum likelihood method [10]. The PML estimator is consistent [11] [12] [13], asymptotically normal [11], and having minimal prediction error [11]. The PML approach has been employed in many models for analysis of complex survey data [14]. There are three possible conditions for using the PML. The first condition was when we made certain assumptions or treatments on the response variable when the independent variable was known [15]. However, this assumption did not apply to all distribution functions of the response variables. The second condition was when we already have the maximum likelihood function but it was difficult to perform numerical/empirical calculations because the computational process was very complex. The last reason was related to the context of the non-nested hypothesis. It involved two interconnected models.

Ordinal logistic regression with sampling weights uses the PML parameter estimation method because we make a certain treatment of the parameter estimation model by adding sampling weights to the response variable when the independent variable was known. [7] described the form of the pseudo-log probability logistic regression model with the sampling weight. The PML in on this study based on [7] research with some modification inline to SUSENAS sampling weight. Let $g(.)$ be a link function such that $\boldsymbol{\pi}$ = **f(x,** $\boldsymbol{\theta}$) where $\boldsymbol{\theta}$ is a p-dimensional column vector for regression coefficients. Then the pseudo-log likelihood form of the logistic regression model with the SUSENAS sampling weight is expressed in Equation (5).

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}}\sum_{m=1}^{\bar{m}}\sum_{o=1}^{o_{khim}} w_{khimo}((\log(\boldsymbol{\pi}_{khimo}))'\boldsymbol{y}_{khimo} + log(\pi_{khimo(D+1)})y_{khimo(D+1)}) \tag{5}$$

For example, the PML parameter estimator is denoted by $\widehat{\boldsymbol{\theta}}$ so that the solution to Equation (5) can be seen in Equation (6) below:

$$\sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}}\sum_{m=1}^{\bar{m}}\sum_{o=1}^{o_{khim}} w_{khimo}\boldsymbol{D}'_{khimo}(diag(\boldsymbol{\pi}_{khimo}) - \boldsymbol{\pi}_{khimo}\boldsymbol{\pi}'_{khimo})^{-1}(\boldsymbol{y}_{khimo} - \boldsymbol{\pi}_{khimo}) = \boldsymbol{0} \tag{6}$$

where $\pi_{khimo(D+1)} = E(y_{khimo(D+1)}|\boldsymbol{x}_{khnmo}) = 1 - \boldsymbol{1}'\boldsymbol{\pi}_{khimo}$, $\boldsymbol{D}_{khnmo}$ is the matrix of partial derivatives of the link function f related to θ.

Iteration is used to get the PML parameter estimator $\widehat{\boldsymbol{\theta}}$ with the initial value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}^{(0)}$. Suppose the first stage represents the stage to estimate $\boldsymbol{\theta}^{(l)}$. The *(l+1)* stage can be expressed by Equation (7).

$$\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + \boldsymbol{Q}^{(l)^{-1}}\boldsymbol{R}^{(l)} \tag{7}$$

where

$$\boldsymbol{Q}^{(l)} = \sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}}\sum_{m=1}^{\bar{m}}\sum_{o=1}^{o_{khim}} w_{khimo}\boldsymbol{D}^{(l)}{}_{khimo}(diag(\boldsymbol{\pi}^{(l)}{}_{khimo}) - \boldsymbol{\pi}^{(l)}{}_{khimo}\boldsymbol{\pi}^{(l)'}{}_{khimo})^{-1}\boldsymbol{D}^{(l)'}{}_{khimo},$$

$$\boldsymbol{R}^{(l)} = \sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}}\sum_{m=1}^{\bar{m}}\sum_{o=1}^{o_{khim}} w_{khimo}\boldsymbol{D}^{(l)}{}_{khimo}(diag(\boldsymbol{\pi}^{(l)}{}_{khimo}) - \boldsymbol{\pi}^{(l)}{}_{khimo}\boldsymbol{\pi}^{(l)'}{}_{khimo})^{-1}(\boldsymbol{y}_{khimo} - \boldsymbol{\pi}^{(l)}{}_{khimo}),$$

$\boldsymbol{D}^{(l)}{}_{khimo}$, $\boldsymbol{\pi}^{(l)}{}_{khimo}$ evaluated on $\boldsymbol{\theta}^{(l)}$.

The iteration process continues until it converges according to the convergent criteria. In general, convergence requires that the reduction of the normalized prediction function is small. Iteration converges at stage *l* if $\frac{g(\boldsymbol{\theta}^{(l)})'\boldsymbol{H}(\boldsymbol{\theta}^{(l)}(\boldsymbol{\theta}^{(l)})}{L(\boldsymbol{\theta}^{(l)}) + 10^{-6}} < \epsilon$, g is the *gradient vector* and $\boldsymbol{H}$ is the *negatif (expected) Hessian Matrix* of the PML function, $\epsilon = 10^{-8}$. Alternatively, the iteration converges when the change in the PML function becomes very small at stage *(l+1)* if $\frac{|L(\boldsymbol{\theta}^{(l+1)}) - L(\boldsymbol{\theta}^{(l)})|}{|L(\boldsymbol{\theta}^{(l)})| + 10^{-6}} < \epsilon$.

3.4 Taylor Linearization

Taylor linearization is a method for estimating variance in surveys using a complex sampling design [16]. Taylor linearization methods are often used to obtain variance estimators and are generally applicable to any sampling design [17]. Taylor linearization straightforward to conduct, can yield consistent and unbiased estimator and standard errors (given the appropriate conditions), and can be performed using a variety of commercially- and freely-available statistical software [18]. [7] describes the variance-covariance matrix estimator in an ordinal logistic regression model with sampling weights using the Taylor Linearization approach. This study adopted the Taylor linearization method compiled by [7] by adjusting the sampling weight of SUSENAS. The variance-covariance matrix estimator is in Equation (8).

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{G}}\hat{\boldsymbol{Q}}^{-1} \tag{8}$$

where

$$\hat{\boldsymbol{Q}} = \sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}}\sum_{m=1}^{\bar{m}}\sum_{o=1}^{o_{khim}} w_{khimo}\hat{\boldsymbol{D}}_{khimo}(\text{diag}\,(\hat{\boldsymbol{\pi}}_{khimo}) - \hat{\boldsymbol{\pi}}_{khimo}\hat{\boldsymbol{\pi}}_{khimo}{}')^{-1}\hat{\boldsymbol{D}}'_{khimo},$$

$$\hat{\boldsymbol{G}} = \frac{O-1}{O-p}\sum_{k=1}^{K}\frac{H(1-f_1)}{H-1}\sum_{h=1}^{H_k}\frac{n_{kh}(1-f_2)}{n_{kh}-1}\sum_{i=1}^{n_{kh}}\frac{\bar{m}(1-f_3)}{\bar{m}-1}\sum_{m=1}^{\bar{m}}(\boldsymbol{e}_{khim.} - \bar{\boldsymbol{e}}_{khi..})'(\boldsymbol{e}_{khim.} - \bar{\boldsymbol{e}}_{khi..}),$$

O is the total number of observation units, $O = \sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}}\sum_{m=1}^{\bar{m}} o_{khim}$,

$\boldsymbol{e}_{khim.} = \sum_{o=1}^{o_{khim}} w_{khimo}\,\hat{\boldsymbol{D}}_{khimo}(diag(\hat{\boldsymbol{\pi}}_{khimo}) - \hat{\boldsymbol{\pi}}_{khimo}\hat{\boldsymbol{\pi}}'_{khimo})^{-1}(\boldsymbol{y}_{khimo} - \hat{\boldsymbol{\pi}}_{khimo})$

$\bar{\boldsymbol{e}}_{khi..} = \frac{1}{\bar{m}}\sum_{m=1}^{\bar{m}}\boldsymbol{e}_{khimo}$, $\hat{\boldsymbol{D}}_{khimo}$ and $\hat{\boldsymbol{\pi}}_{khimo}$ evaluated on $\hat{\boldsymbol{\theta}}$.

3.5 Model Evaluation

The test data is used to obtain the accuracy of the prediction results. The accuracy value is calculated from confusion matrix [19]. The proportion of training and testing data is subjective depending on the researcher on the condition that the percentage of training data is greater than the testing data [20].

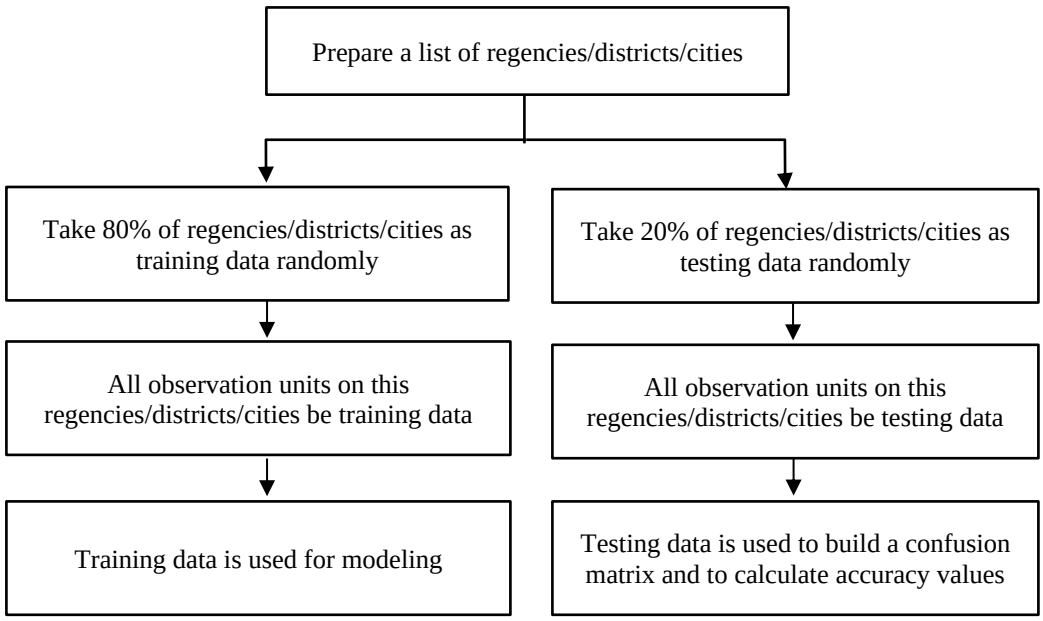

Figure 1. The Flow of the Distribution of Training Data and Testing Data

One of the challenges in compiling a classification accuracy table is dividing the data into training data and testing data by taking into account the sampling weight. The process to divide the data cannot be done directly by randomizing the sample of observation unit because it can affect the weight of the sample that has been compiled by BPS. Figure 1 describes the flow of the distribution of training data and testing data. From a list of regencies/districts/cities in Indonesia, 80% of regencies/districts/cities were randomly selected

as training data and 20% as testing data. All the observation units in the regencies/districts/cities selected as training data are used as training data. All the observation units in the regencies/districts/cities selected as testing data are used for testing data. Training data are used to construct an ordinal logistic regression model with sampling weights. Testing data is used to compile a classification accuracy table and calculate the accuracy value. This value is used to evaluate the model.

3.6 Numerical Example

To make it easier to understand the results of the analysis, in this paper ordinal logistic regression with sampling weight will be directly applied to the March 2020 SUSENAS data. The observation unit is the elderly in Indonesia. The elderly population is someone aged 60 years and over [21]. The total sample used is 121,961 elderly spread across 514 regencies/districts/cities in Indonesia.

From 514 regencies/districts/cities in Indonesia, 80% of regencies/districts/cities (411 regencies/districts/cities) were randomly selected as training data and 20% as testing data (103 regencies/districts/cities). All the observation units in the regencies/districts/cities selected as training data are used as training data to construct the model (98.612 elderly). All the observation units in the regencies/districts/cities selected as testing data are used for testing data to compile a classification accuracy table and calculate the accuracy value (23.349 elderly).

The response variable used is the health status of the elderly. The response variables consisted of three categories, namely 1 = poor health status (experiencing health complaints and not interfering with daily activities), 2 = moderate health status (experiencing health complaints but not interfering with daily activities), and 3 = good health status (no health complaints). There are nine explanatory variables used, namely education level (1 = did not pass elementary school, 2 = pass elementary school or equivalent, 3 = pass junior high school or equivalent, 4 = pass senior high school or equivalent, 5 = college graduated), gender (1 = male, 2 = female), marital status (1 = married, 2 = not married/divorced), work status (1 = does not work, 2 = work), type of area of residence (1 = rural, 2 = urban), smoking habit (1 = do not smoke, 2 = light smoker, 3 = moderate smoker, 4 = heavy smoker), household food insecurity status (1 = not food insecure, 2 = food insecurity), number of household members, and average monthly expenditure per capita (1 = less than 1 million rupiahs, 2 = 1 to 2 million rupiahs, 3 = more than 2 million rupiahs).

The results of the G test show that the ordinal logistic regression model with sampling weights and without sampling weights is significant at alpha 5% (p-value = 0.0000). From Table 2, it can be seen that there are significant differences in the p-value for the variables of marital status and the type of area of residence. In the ordinal logistic regression model without sampling weight, the variables of marital status and type of area of residence significantly affect the health status of the elderly (p-value < alpha 5%). In the ordinal logistic regression model with sampling weights, these variables do not significantly affect the health status of the elderly (p-value > alpha 5%). The standard error for ordinal logistic regression without sampling weight is smaller than ordinal logistic regression with sampling weight.

Table 2. Standard Error and P-Value Ordinal Logistic Regression without and with Sampling Weight by Variable

Variable	Ordinal Logistic Regression without Sampling Weight		Ordinal Logistic Regression with Sampling Weight	
	Standard Error	P-Value	Standard Error	P-Value
Education level5	0.0320	0.0000	0.0494	0.0000
Education level2	0.0144	0.0000	0.0236	0.0001
Education level3	0.0238	0.0000	0.0371	0.0000
Education level4	0.0248	0.0000	0.0379	0.0000
Gender2	0.0158	0.0000	0.0226	0.0000
Marital status2	0.0140	**0.0008**	0.0209	**0.0756**
Work status2	0.0139	0.0000	0.0213	0.0000
Type of area of residence2	0.0133	**0.0192**	0.0240	**0.4935**

Variable	Ordinal Logistic Regression without Sampling Weight		Ordinal Logistic Regression with Sampling Weight	
	Standard Error	P-Value	Standard Error	P-Value
Smoking habit1	0.0417	0.0971	0.0652	0.8664
Smoking habit2	0.0451	0.0097	0.0695	0.0033
Smoking habit3	0.0460	0.0000	0.0708	0.0000
Household food insecurity status2	0.0151	0.0000	0.0267	0.0000
Number of household members	0.0033	0.0000	0.0054	0.0000
Average monthly expenditure per capita1	0.0223	0.1505	0.0356	0.4866
Average monthly expenditure per capita2	0.0218	0.0064	0.0344	0.0088
Intercept 1|2	0.0487	0.0000	0.0772	0.0000
Intercept 2|3	0.0487	0.0000	0.0771	0.0000

Table 3 shows that the percentage of elderly who have good, moderate, and poor health status for those who are married and not married/divorced is not too different. The same pattern also occurs in the variable type of residential area, both those living in urban and rural areas. This shows that the results of the ordinal logistic regression model with sampling weights in Table 2 better describe the actual condition of the health status of the elderly than the ordinal logistic regression model without sampling weights.

Table 3. Percentage of Elderly by Marital Status, Type of Area of Residence, and Health Status

Variable	Health Status			Total
	Poor	Moderate	Good	
Marital Status				
Married	23.14	23.25	53.61	100
Not married/ divorced	26.15	24.44	49.41	100
Type of area of residence				
Urban	22.81	25.29	51.90	100
Rural	26.03	21.91	52.06	100

The accuracy values of the ordinal logistic regression model with sampling weights and without sampling weights are the same (53%). However, Table 4 and Table 5 show that the ordinal logistic regression model with sampling weights more capable to predict poor and moderate health status of the elderly than the ordinal logistic regression model without sampling weights.

Table 4. Confusion Matrix Ordinal Logistic Regression without Sampling Weight

Health Status (Prediction Result)	Health Status (Actual Condition)		
	Poor	Moderate	Good
Poor	936	599	771
Moderate	0	0	0
Good	4760	4848	11435

Table 5. Confusion Matrix Ordinal Logistic Regression with Sampling Weight

Health Status (Prediction Result)	Health Status (Actual Condition)		
	Poor	Moderate	Good
Poor	1074	716	938
Moderate	40	28	54
Good	4582	4703	11214

4. CONCLUSIONS

Ordinal logistic regression with sampling weights uses the Pseudo Maximum Likelihood parameter estimation method because we make a certain treatment of the parameter estimation model by adding sampling weights to the response variable when the independent variable was known. The method for estimating variance used Taylor linearization. The testing data is used to calculate the accuracy of the prediction results. The process to divide the data cannot be done directly by randomizing the sample of the observation unit because it can affect the sampling weight that has been compiled by BPS. Distribution of training data and testing data is selecting 80% of regencies/districts/cities randomly as training data and 20% as test data. All observation units in selected regencies/districts/cities as training data are used as training data while others as testing data. The results of the numerical example indicate that the ordinal logistic regression model with sampling weights is more representative of the population and more capable to predict minority categories of the response variable (poor and moderate health status) than without using sampling weights.

AKNOWLEDGEMENT

The authors expressed our gratitude to BPS-Statistics Indonesia for funding support in this research.

REFERENCES

[1] M. W. Fagerland and D. W. Hosmer, "How to test for goodness of fit in ordinal logistic regression models," *Stata Journal*, vol. 17, no. 3, pp. 668-686, 2017.

[2] K. J. Archer and S. Lemeshow, "Goodness of Fit Test for a Logistic Regression Model Fitted Using Survey Sample Data," *The Stata Journal*, vol. 6, no. 1, pp. 97-105, 2006.

[3] Badan Pusat Statistik, Pedoman Kepala BPS Provinsi, Kepala Bidang Statistik Sosial, dan Kepala BPS Kabupaten/Kota: SUSENAS Maret 2020, Jakarta: Badan Pusat Statistik, 2019.

[4] M. A. Ciol, J. M. Hoffman, B. J. Dudgeon, A. Shumway-Cook, K. M. Yorkston and L. Chan, "Understanding the use of weights in the analysis of data from multistage surveys," *Arch Phys Med Rehabil*, vol. 87, no. 2, pp. 299-303, 2006.

[5] S. R. Cassy, I. Natario and M. R. O. Martins, "Logistic Regression Modelling for Complex Survey Data with an Application for Bed Net Use in Mozambique," *Open Journal of Statistics*, vol. 6, no. 5, pp. 898-907, 2016.

[6] K. S. Barasa and C. Muchwanju, "Incorporating Survey Weights into Binary and Multinomial Logistic Regression Models," *Science Journal of Applied Mathematics and Statistics*, vol. 3, no. 6, pp. 243-249, 2015.

[7] B. Anthony, "Performing Logistic Regression on Survey Data with the New SURVEYLOGISTIC Procedure," in *The Twenty-Seventh Annual SAS®Users Group International Conference*, Orlando, 2002.

[8] J. Jajang, N. Nurhayati and S. J. Mufida, "Ordinal Logistic Regression Model and Classification Tree on Ordinal Response Data," *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, no. 1, pp. 075-082, 2022.

[9] G. Solomon and L. Weissfeld, "Pseudo maximum likelihood approach for the analysis of multivariate left-censored longitudinal data," *Stat Med.*, vol. 36, no. 1, pp. 81-91, 2017.

[10] A. Guolo, "Pseudo-Likelihood inference for regression models with misclassified and mismeasured variables," *Statistica*

Sinica, vol. 21, pp. 1639-1663, 2011.

[11] M. Abdalmoaty and H. Hjalmarsson, "Simulated Pseudo Maximum Likelihood Identification of Nonlinear Models," in *The 20th IFAC World Congress*, Elsevier, 2017.

[12] G. Fiorentini and E. Sentana, "Consistent non-Gaussian pseudo maximum likelihood estimators," *Journal of Econometrics*, vol. 213, no. 2, pp. 321-358, 2019.

[13] C. Gourieroux, A. Monfort and E. M. Renault, "Consistent Pseudo-Maximum Likelihood Estimators," *Annals of Economics and Statistics*, vol. 125/126, pp. 187-218, 2017.

[14] J. Wang, "The Pseudo Maximum Likelihood Estimator for Quantiles of Survey Variables Get access Arrow," *Journal of Survey Statistics and Methodology*, vol. 9, no. 1, p. 185–201, 2021.

[15] M. Denuit, D. Hainaut and J. Trufin, Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions, Switzerland: Springer, 2019.

[16] S. L. Lohr, Sampling: Design and Analysis, Second Edition, Boston: Brooks/Cole, 2010.

[17] A. Demnati and J. N. K. Rao, "Linearization variance estimators for model parameters from complex survey data," *Survey Methodology*, vol. 36, no. 2, pp. 193-201, 2010.

[18] F. L. Haung, "Analyzing Group Level Effects with Clustered Data Using Taylor Series Linearization," *Practical Assessment, Research, and Evaluation*, vol. 19, no. 1, pp. 1-9, 2014.

[19] R. Prasetya and A. Ridwan, "Data Mining Application on Weather Prediction Using Classification Tree, Naïve Bayes and K-Nearest Neighbor Algorithm with Model Testing of Supervised Learning Probabilistic Brier Score, Confusion Matrix and ROC," *Journal of Applied Communication and Information Technologies*, vol. 4, no. 2, pp. 25-33, 2019.

[20] S. Aisyah, S. Wahyuningsih and F. Amijaya, "Peramalan Jumlah Titik Panas Provinsi Kalimantan Timur Menggunakan Metode Radial Basis Function Neural Network," *Jambura Journal of Probability and Statistics*, vol. 2, no. 2, pp. 64-74, 2021.

[21] United Nations Population Fund (UNFPA) and HelpAge International, "Ageing in the Twenty-First Century: A Celebration and A Challenge," UNFPA and HelpAge International, New York and London, 2012.