Application of Logistic Regression Modeling for Complex Survey Data on Education Continuity of Poor Households Children

R Salam[1,4]**, A Adji** [2,3]

[1] Polytechnic of Statistic STIS, Jakarta, Indonesia
[2] BPS Statistics Indonesia, Jakarta, Indonesia
[3] The National Team for The Acceleration of Poverty Reduction, Jakarta, Indonesia

*Corresponding author's e-mail: rudisalam@stis.ac.id

Abstract. Many population-based surveys such as the National Socio-Economic Survey (Susenas) are built with complex sampling assumptions, namely probabilistic, stratified, and multistage sampling, with unequal weights for each observation. This complex design must be taken into account in order to have reliable results when doing modeling. The model that is often used when using survey data is logistic regression. The purpose of this study is to determine a logistic regression model with a complex sample design, and to show how it is estimated using a package survey from the R software. As an illustration, the 2019 Susenas data of East Java Province will be used as an application to correct the influence of the sample design in estimating risk factors related to the chances of children 7-18 years old in poor households continuing their education. The results show that the variables of gender and mother's education significantly affect the continuity of the education of children 7-18 years old in poor households.

1. Introduction

Research using survey data such as the National Socio-Economic Survey (Susenas) by the BPS-Statistics Indonesia, Indonesian Demographic and Health Survey (IDHS) by the National Population and Family Planning Board (BKKBN) has become a common thing for researchers today because of the easy access to data from the survey results. Researchers in analyzing the survey data usually have research questions to be answered and appropriate analytical methods that will be used if the data used is selected using simple random sampling. However, there are problems in the analysis when the data used is collected through a complex survey design involving stratification, clustering, multistage sampling, and determining unequal opportunities in sample selection. Suppose a sample is taken using a clustering method, observations from the same group will be correlated and in order to obtain an unbiased estimator, it is necessary to weight the sample to adjust the influence of the cluster. Ignoring the sampling method in analyzing data can lead to inaccurate results (Cassy, et al. 2016). Some authors evaluate the adverse consequences if they ignore the sampling scheme in statistical analysis (Pessoa and Silva, 1998). That means that to make valid conclusions for a population where the sample comes from, appropriate statistical methods are needed to analyze complex survey data.

One survey that uses a complex sample design is Susenas. Susenas includes data relating to socio-economic conditions of the community including health, education, fertility, family planning, housing conditions and other socio-economic conditions. Susenas data presents a lot of data in the form of categories. One of the statistical methods commonly used in predicting results from categorical data

from one set of covariates is logistic regression model. Logistic regression model is a member of the generalized linear model (GLM) class and is an appropriate model for studying the relationship between binary response variable Y, which represent success ($Y=1$) or failure ($Y=0$), and a collection of covariates $\mathbf{x}=\left(x_1,\ldots,x_p\right)^T$. Assuming the response variable Y has a Bernoulli distribution, the model can be written as follows:

$$g\left(x\right)=\ln\left(\frac{\pi\left(x\right)}{1-\pi\left(x\right)}\right)=\beta_0+\beta_1x_1+\ldots+\beta_px_p \tag{1}$$

or equivalently

$$\pi\left(x\right)=\frac{\exp\left(\beta_0+\beta_1x_1+\ldots+\beta_px_p\right)}{1+\exp\left(\beta_0+\beta_1x_1+\ldots+\beta_px_p\right)} \tag{2}$$

where $\beta_0,\beta_1,\ldots,\beta_p$ is an unknown parameter that will be estimated and $\pi\left(x\right)=P\left(Y=1|\,x\right)$ is a probability of success.

The model parameters are estimated by the maximum likelihood method which assumes that the observations are mutually independent and identical distributed. However, under a complex sample design, the independent assumption between observations is usually not fulfilled. The parameters estimation by the maximum likelihood method can produce incorrect standard error values and consequently the hypothesis test results also become incorrect. Therefore, it is necessary to adjust a standard logistic regression method that considers complex sampling designs in order to produce valid conclusions (Pessoa & Silva, 1998; Hosmer & Lemeshow, 2000; Lee & Forthofer, 2006).

Some studies in health science analyze data derived from complex sampling designs and use various types of software (Chinomona and Mwambi, 2015; Baume and Franca-Koh, 2011; Mutuku et. al., 2013; Zango et. al., 2013). This paper focuses on presenting a logistic regression model framework for complex sampling design using R.

As an application to the logistic regression model for this complex survey data, category data will be used whether the education of children aged 17-18 years in poor households continues or not from Susenas in March 2019. Susenas data is collected through a complex survey design that involves stratification, clustering, multistage sampling, and determination of unequal opportunities in sample selection so that when you want to model data from Susenas, you should choose a method that has considered the sampling design as has been implemented in the survey package on the R software (Lumley, 2010; Lumley, 2015). Therefore, the aim of this study is to apply a logistic regression method that has considered the influence of the sampling design to model the educational continuity status of children aged 7-18 years in poor households and the variables that influence it in East Java in 2019.

2. Logistic regression with complex survey

Hosmer (2002) and Lee and Forthofer (2006) state that ordinary logistic regression model is not appropriate to use if data are obtained from complex sampling designs. Suppose that a finite population $U=\left\{1,2,\cdots,N\right\}$ is divided into $h=1,2,\cdots H$ strata. Each strata is further divided into $j=1,2,\ldots,n_h$ primary sampling units (PSU), each of which is constituted by $i=1,2,\ldots,n_{hj}$ secondary sampling unit (SSU), each comprehending n_{hji} elements. If it is also assumed that the observed data consists of n'_{hj} SSU selected from n'_h PSU in the stratum h. The number of overall observations is

$$n=\sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}n_{hji} \tag{3}$$

Each sampling unit has a corresponding sampling weight which is inverse of its probabilities in the sample, denoted by:

$$w_{hjik} = \frac{1}{\pi_{hjik}} \text{ for the } hjik\text{-th unit.} \tag{4}$$

In addition, suppose that Y_{hjik} denote the binary response variable, $\mathbf{x}_{hjik}$ denote the covariate matrix and $\boldsymbol{\beta}$ denoted the regression coefficient. Thus in general, the survey logistic regression model is given by

$$\log it\left\{P\left(Y_{hjik}=1|\,x_{hjik}\right)\right\} = \ln\left\{\frac{P\left(Y_{hjik}=1|\,x_{hjik}\right)}{1-P\left(Y_{hjik}=1|\,x_{hjik}\right)}\right\} = \mathbf{x}_{hjik}^{T}\boldsymbol{\beta} \tag{5}$$

So under a complex sampling design, the parameters $\boldsymbol{\beta}$ of the logistic regression model are estimated by the maximum pseudo-likelihood method are also known as weighted maximum likelihood which combines sampling design and sampling weights that are different in estimation of the $\boldsymbol{\beta}$ (Hosmer and Lemeshow, 2000; Lee and Forthofer, 2006 Archer, Lemeshow and Hosmer, 2007; Lumley, 2004). The main idea of this method is to define a function that approaches the likelihood function of finite population samples with the likelihood function formed by the observation sample and the known sampling weight (Hosmer and Lemeshow, 2000; Lee and Forthofer, 2006; Archer, Lemeshow and Hosmer, 2007; Lumley, 2004). In this case, the pseudo-log-likelihood is given by

$$\begin{aligned} l_p(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}\sum_{k} w_{hjik}\Big\{ & y_{hjik}\ln\left[P\left(Y_{hjik}=1|\,\mathbf{x}_{hjik}\right)\right] \\ & +\left(1-Y_{hjik}\right)\ln\left[1-P\left(Y_{hjik}=1|\,\mathbf{x}_{hjik}\right)\right]\Big\} \end{aligned} \tag{6}$$

where w_{hjik} is the weight of the observation $hjik$. The maximum pseudo-likelihood estimator of $\boldsymbol{\beta}$ is obtained by deriving the pseudo-log likelihood function in order to $\boldsymbol{\beta}$ and equal is to zero, $(\boldsymbol{\beta}) = \frac{d}{d\beta} l_p(\boldsymbol{\beta}) = 0$.

Under a complex sampling design, there is no form that can directly calculate variance estimators. Thus, to obtain variance estimators with maximum pseudo-likelihood, methods such as Taylor linearization (also known as delta method), Jackknife and bootstrap replication (Hosmer and Lemeshow, 2000; Lee and Forthofer, 2006; Lumley, 2004) are used. In this paper, R is used with a package survey where this package applies methods such as Taylor linearization (Lumley, 2010). The variance estimator from which this method is produced is

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{S}\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1} \tag{7}$$

where $\mathbf{X}$ is the covariate matrix, $\mathbf{D} = \mathbf{W}\mathbf{V}$ is the diagonal nxn matrix with elements $w_{hjik} \times \hat{P}\left(Y_{hjik}=1|\,\mathbf{x}_{hjik}\right)\left[1-\hat{P}\left(Y_{hjik}=1|\,\mathbf{x}_{hjik}\right)\right]$, and $\mathbf{S}$ is a pooled estimator within-stratum of the covariance matrix where the value is

$$\mathbf{S} = \sum_{j=1}^{H}\left(1-f_h\right)\frac{n'_h}{n'_h-1}\sum_{j=1}^{n'_h}\left(\mathbf{z}_{hj..}-\tilde{\mathbf{z}}_{h...}\right)\left(\mathbf{z}_{hj..}-\tilde{\mathbf{z}}_{h...}\right)' \tag{8}$$

where $\mathbf{z}_{hjik} = w_{hjik} \times \hat{P}\left(Y_{hjik}=1|\,\mathbf{x}_{hjik}\right)\left[1-\hat{P}\left(Y_{hjik}=1|\,\mathbf{x}_{hjik}\right)\right]$, being the sum for all the n'_{hj} sampled units in PSU j in the stratum h given as $\mathbf{z}_{hj..} = \sum_{i=1}^{n'_{hj}}\mathbf{z}_{hjik}$ and specific mean in the stratum as

$\bar{\mathbf{z}}_h = \frac{1}{n'_h}\sum_{j=1}^{n'_h}\mathbf{z}_{hj..}$. The correction factor is given by $(1-f_h)$, where $f_h = \frac{n'_h}{n_h}$ is the ratio of the number of PSU observed by the total number of PSU in the stratum h.

Hypothesis testing for the significance of the regression coefficients and the test for the goodness of model fit also need to be modified to incorporate the sampling design and different weighting of observations. Evaluations of covariate contributions are now made with an adjusted Wald test (Lee and Forthofer, 2006), with test statistics as follows (Hosmer and Lemeshow, 2000; Lee and Forthofer, 2006; Lumley and Scott, 2014)

$$F = \frac{s-p+1}{sp}W$$
$$W = \hat{\boldsymbol{\beta}}^T\left[Var\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\hat{\boldsymbol{\beta}} \quad (9)$$

where $s = \sum_{h=1}^{H} n'_h - H$ is the total number of selected PSU minus the number of strata and p is the number of covariates. The F statistics above is distributed as a F-distribution with p and $(s-p+1)$ degrees of freedom, so $p-value = P\left[F\left(p, s-p+1\right) \geq F\right]$.

Also, in order to obtain valid inferences using this type of design, we introduced Pearson's test statistic, such as the Rao-Scott adjustments. Alternatively another test statistic can be used that has included a sampling plan such as the adjusted Wald statistics (Hosmer and Lemeshow, 2000; Lee and Forthofer, 2006). We also use Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare models (Archer, Lemeshow, and Hosmer, 2007; Lumley and Scott, 2015), and likelihoods from Lumley and Scott (2014) to measure goodness of fit which considers a complex sampling frame.

3. Methods

Susenas is one of the surveys held regularly by BPS every year. Susenas collects data relating to the socio-economic conditions of the community, which includes the conditions of health, education, fertility, family planning, housing and other socioeconomic conditions. Since 2015, Susenas has been implemented twice a year, in March for district/city estimates, and in September for provincial estimates. The total number of households in the 2015 Susenas sample is 300,000 households for district/city estimates in March and 75,000 households for provincial estimates in September.

The 2019 Susenas sample was selected using the two stages one phase stratified sampling method where the first stage was to choose 25% population census block with Proportional to Size (PPS) probability, with the size of the number of SP2010 households in each stratum. The second stage is to choose a number of census blocks according to systematic allocation in each urban/rural strata per district/city per welfare stratum. The third stage is to select 10 households that are updated by systematic sampling with implicit stratification according to the highest education that is completed by the head of household.

Table 1. Susenas sampling scheme.

Phase	Unit	Unit number of stratum h		Sampling methods	Probability of sampling selection	Sampling fraction
		Population	Sample			
1	Census Block	N_h	n'_h	PPS WR	Z_{hi}/Z_h	$n'_h\frac{Z_{hi}}{Z_h}$
		n'_h	n_h	Systematic	$\frac{1}{n'_h}$	$\frac{n_h}{n'_h}$

Phase	Unit	Unit number of stratum h		Sampling methods	Probability of sampling selection	Sampling fraction
		Population	Sample			
2	Households	M_{hi}^{up}	$\bar{m}$	Systematic	$\frac{1}{M_{hi}^{up}}$	$\frac{\bar{m}}{M_{hi}^{up}}$

In this study, the unit to be analyzed is children aged 7-18 years in poor households in East Java in 2019 with a total sample of 1451 children.

The method used in this paper is logistic regression with survey data applied to identify the factors that influence children 7-18 in poor households whether their education continues or not. Table 2 shows the variables used in this study.

Table 2. List of response and explanatory variables.

Variables	Remarks
Response variable:	
- Continuity status	0 = Not continue, 1 = Continue
Explanatory variable:	
- Gender	0 = Female, 1 = Male
- Mother's education	0 = Secondary below, 1 = Highschool above
- Number of household member	0 = ≤ 3, 1 = > 3

To find a fit model with data used multiple logistic regression models. The last model used only contains real variables related to women's EBI status to a significance level of 5%. After that, the odds ratio (OR) is calculated. Because of the complex nature of data sampling, all analyzes are carried out using survey packages from R software (Lumley, 2015) where all design features such as stratification, grouping and weighting are explicitly recorded using the svydesign function. To describe the model, by determining the predictor and the connecting function used, the svyglm function is used. To test the goodness of the model follow what has been done and explained in Lumley (2010).

4. Results

4.1. Descriptive

The sample of this study was 1451 children aged 7-18 years from poor households in East Java in 2019. Of the 1451 children, there were still around 9.6% whose education did not continue. Of the total 9.6%, 1% of those who have not/never been in school are and 8.5% of those who are no longer in school. In terms of gender characteristics, women are less likely to discontinue than men and men are more likely to continue their education than women. Table 2 also shows the relationship between several household characteristics and the sustainability of children's education in poor households. From the characteristics of the mother's education, the percentage of children whose education did not continue was greater than that of mothers with lower education (97.1%). Likewise with the characteristics of the number of household members, the greater the number of household members, the greater the percentage of children whose education does not continue will be even greater.

Table 3. Percentage of education continuity of poor household children aged 7-18 by individual and household characteristics

Characteristics	Education continuity		Total
	Not continue	Continue	
Gender			
- Male	49,6	52,2	52,0
- Female	50,4	47,8	48,0
Mother's education			
- Secondary and below	97,1	91,8	92,3
- Highschool above	2,9	8,2	7,7
Household member			
- <=3	13,7	8,9	9,4
- >3	86,3	91,1	90,6

4.2. Logistic regression estimate

In order to be able to explore the effect of these variables on the sustainability of children's education, an estimation of a binary logistic regression model has been carried out in which the dependent variable is the sustainability of education, namely "yes" if education continues and "no" if education does not continue. This study has a limited sample used, namely children aged 7-18 years in poor households.

In this paper all covariates in Table 2 are included in the analysis. Table 3 shows the results of logistic regression estimation with complex sampling. The results of the processing found that the variables that significantly affected the sustainability of children's education in poor households were gender and mother's education. Taking into account gender, male children will have a tendency to remain in school 1.03 times compared to female children. With an odds ratio of around one, it can be said that there is no significant difference between men and women in poor households in terms of education sustainability.

When viewed from the characteristics of the household, namely maternal education, children from poor households with a mother's education of high school and above compared to children with a mother's education of junior high school have a tendency of 1.09 times for the continuation of their child's education. Apart from mother's education, household characteristics are also represented by the number of household members. The results show that the tendency of children with more than 3 household members to continue their education is 1.06 times compared to children with fewer household members (less than equal to 3).

Table 4. Logistic regression estimate with complex sampling.

Variables	Coefficients	Standard error	p-value	OR
Intersep	0.8996	0.0415	0.0000	2.4587
Jenis kelamin	0.0342	0.0182	0.0603	1.0348
Pendidikan ibu	-0.0845	0.0194	0.0000	0.9190
Jumlah art	0.0629	0.0399	0.1149	1.0650

5. Conclusions

Complex sampling frames are already widely used for population-based surveys such as Susenas. However, the complexity of the methodology because it involves gradual stratification, grouping, and sampling is still not well understood by observers in various fields of applied science. From this paper it can be shown that it is possible to obtain reliable results and more efficient estimators by using an appropriate method that can correct for the influence of the sample design. Furthermore, this paper and the availability of open source software such as R should encourage scientists to use more easily accessible survey data.

From the applied side, the results of this study indicate that the variables that influence the sustainability of children's education in poor households are gender and mother's education where mother's education has a very significant effect. In order to get optimal results in efforts to eradicate poverty, it is better if poverty reduction programs related to education are carried out by taking into account the needs of each poor household, for example the number of children who go to school, especially girls so that later they can become mothers who pay attention to their children's education.

References

[1] K. Archer, S. Lemeshow and D. Hosmer, "Goodness-of-Fit Tests for Logistic Regression Models When Data Are Collected Using a Complex Sampling Design," Computational Statistics and Data Analysis, 51, pp. 4450-4464, 2007.

[2] C. Baume and A. Franca-Koh, "Predictors of Mosquito Net Use in Ghana," Malaria Journal, 10, p. 265, 2011.

[3] S. Cassy, I. Natário and M. Martins, "Logistic Regression Modelling for Complex Survey Data with an Application for Bed Net Use in Mozambique," Open Journal of Statistics, 6, pp. 898-907, 2016.

[4] D. Hosmer and S. Lemeshow, Applied Logistic Regression. 2nd Edition, Wiley Series in Probability and Statistics, 2000.

[5] A. Irawati, "Inisiasi menyusui dini dan factor determinannya pada anak balita di Indonesia: analisis data sekunder survey demografi dan kesehatan Indonesia Tahun 2007," PGM 2010, vol. 33(1), pp. 1-13, 2010.

[6] E. Lee and R. Forthofer, Analyzing Complex Survey Data. 2nd Edition, Sage, Thousand Oaks., 2006.

[7] T. Lumley, "Analysis of Complex Survey Samples," Journal of Statistical Software, vol. 9, pp. 1-19, 2004.

[8] T. Lumley, Complex Surveys: A Guide to Analysis Using R, Hoboken, Washington: John Wiley and Sons, 2010.

[9] T. Lumley and A. Scott, "Tests for Regression Models Fitted to Survey Data," Australian and New Zealand Journal of Statistics, vol. 56, pp. 1-14, 2014.

[10] T. Lumley, Survey: Analysis of Complex Survey Samples, R Package Version 3.31-0, 2015.

[11] F. Mutuku and e. al., "Physical Condition and Maintenance of Mosquito Bed Nets in Kwale County, Coastal Kenya," Malaria Journal, vol. 12, p. 46, 2013.

[12] D. Pessoa and P. Silva, Análise de dados amostrais complexos, São Paulo, Brasil: Associação Brasileira de Estatística, 1998.

[13] A. Chinomona and H. Mwambi, "Estimating HIV Prevalence in Zimbabwe Using Population-Based Survey Data," PloS ONE, Vols. 10, e0140896., 2015.

[14] A. Zango and e. al., "Determinants of Prevalent HIV Infection and Late HIV Diagnosis among Young Women with Two or More Sexual Partners in Beira, Mozambique," PLoS ONE, Vols. 8, e63427., 2013.