

MULTIVARIATE REGRESSION IN COMPLEX SURVEY DESIGN

G. M. Oyeyemi, A. A. Adewara

Department of Statistics
University of Ilorin
Ilorin, Nigeria

S. O. Ige

Department of Mathematics & Statistics
Federal Polytechnic Ado-Ekiti
Ado-Ekiti, Nigeria

B. A. Ibraheem

Department of Statistics
Kwara State Polytechnic
Ilorin, Nigeria

ABSTRACT

In multivariate regression analysis of complex survey design, adjustment for both stratification and clustering variable depends on whether the variables are exogenous or endogenous in the model. Adjustments are found to be necessary if the variables are exogenous in the models, especially adjustment for clustering is found to be more important than that of stratification.

KEYWORDS: Adjustment, Complex Survey, Endogenous, Exogenous, Cluster, Strata

INTRODUCTION

In most of demographic and health surveys, the basic descriptive statistics is given by the bivariate relationship between a health variable and some health indicators known as socio-economic status (SES). More importantly, one may want to test the existence of causal relationship between a health variable and the socio-economic status (SES) variables and to examine the nature of the causality (Fergusen and Corey, 1990). These tasks will require moving from bivariate to multiple or multivariate regression analysis. In this paper, we examine some issues that generally deserve consideration when undertaking multivariate analysis in health survey, especially when dealing with a complex survey design.

1. COMPLEX SURVEY HEALTH SURVEY DESIGN

Demographic and health surveys used for analysis of health sector in developing countries have complex sample designs. Typically, there is simple random sampling at some level or levels but there might be separate sampling from population subgroups known as strata. Also there is possibility of groups of observations, otherwise known as clusters, which may not be sampled independently and there may be over sampling of certain groups. The combination of these different sampling schemes constitutes what is known as a complex design. In the following sections we discuss the components of complex survey design and the needs for standard errors adjustment of the estimates.

2. STRATIFICATION

Stratification is a method of using auxiliary variable to increase the precision of the estimate of a population characteristic (Cochran, 1977; Okafor, 2002; Rajj & Chandhok, 1999). This design method is typically employed in household surveys undertaken in developing countries. Nigerian Demographic and Health Survey (NDHS) use geopolitical zones (regions) as the stratification variable in most of its surveys. A random sample, of predetermined size, is then selected independently from each of these strata. The sample accounted for by each stratum may or may not correspond to population proportion. There is equal allocation, proportional allocation, optimum allocation among others depending on the design and situation.

In case the sample proportions do not correspond to population proportions, the overall sample is not representative of the population and the issue of sample weight arises. If the population means differ across the strata, predetermination of strata sample sizes reduces the sampling variance of the estimators of the means (Fergusen and Corey, 1990). Consequently, standard errors of estimates of population means and some other statistics should be adjusted. Also, adjustment is not necessary in regression analysis and in a wide variety of other multivariate modelling approaches provided stratification variable is exogenous within the model (Wooldridge, 2001; 2005). Ordinary Least Square (OLS) estimation is found to be consistent and efficient and the usual standard errors are valid in such case. But if the stratification is based on an endogenous variable then the standard errors should be adjusted (Wooldridge, 2001).

Therefore, the need to adjust standard errors for stratification is situation specific. In practice, adjusting for stratification may inflate the standard errors. But with survey data, standard errors robust to heteroscedasticity and clustering will be required (Hao *et al*, 2004). The magnitude of adjustment for stratification is usually modest and normally downward. A conservative strategy is not to make adjustment. If the stratification variable is exogenous, there is need for adjustment and if endogenous, any adjustment will normally increase statistical significance.

3. CLUSTERING

In most surveys, especially a large and complex survey, there is no sampling frame or list of households or dwelling units from which to select a sample. Therefore it is not possible or feasible to draw a sample directly from the population. In order to overcome this problem, groups of elements called clusters are formed by pulling together elements which are physically close to each other (Cochran, 1977 and Okafor, 2002). A sample of these cluster units is then selected from the total number of clusters by an appropriate sampling scheme. This method is known as cluster sampling and a sampling unit in this case (cluster) contains more than one unit.

Cluster sampling in health survey has two stages or more sampling process. In the first stage, groups known as clusters of households are randomly sampled from either the population or strata. Typically, these clusters are villages, hamlets or neighbourhoods of towns and cities. In the second stage, households are randomly sampled from each of the selected clusters. An important distinction of cluster sampling from stratifying sampling is that strata are selected deterministically, whereas clusters are selected randomly (Owen *et al*, 2007). Another difference is that strata are typically few in number and contain many observations, whereas clusters are large in number but contain relatively few observations.

As a result of this design, observations are expected not to be independent within clusters but may be independent across clusters. There is likely to be more homogeneity within cluster than across clusters. The adjustment in standard errors of the estimates like stratification sampling depends on whether the cluster variable has endogenous or exogenous effects on the model.

4. RANDOM EFFECTS MODEL

A cluster or stratification variable is exogenous in a model if it has random effects on the model. Consider the following model:

$$y_{ic} = X_{ic}\beta + \lambda_c + \varepsilon_{ic} \quad (i)$$

$$E[\varepsilon_{ic}/X_{ic}, \lambda_c] = [\varepsilon_{ic}] = 0$$

where i and c are the household (individual) and cluster (or stratification) variables respectively; X_{ic} is a vector of regressors; λ_c is cluster (or stratification) effects; and ε_{ic} are the error terms. If we assume that the cluster (stratification) effects are independent of the regressors; $E[\lambda_c / X_c] = E[\lambda_c]$, Then the composite error is; $u_{ic} = \lambda_c + \varepsilon_{ic}$. This model is known as random effects model.

A conventional point estimates like ordinary least square (OLS) is consistent but inefficiency arises from the fact that the cluster (or stratification) induced correlation in the

G. M. OYEYEMI, A. A. ADEWARA, S. O. IGE, B. A. IBRAHEEM

composite errors, which in addition, requires adjustment of the standard errors. One option is to accept inefficiency and simply adjust the standard errors. An alternative strategy is to presume consistency by estimating the within cluster or strata correlation and taking account of this in estimation of the model parameters. In the linear case, an analyst would use generalized least square (GLS). A Lagrange multiplier test can be used to test the null that the cluster (or stratification) effects are not significant and that ordinary least square (OLS) is efficient (Wooldridge, 2005).

5. FIXED EFFECTS MODEL

Using the model in equation (i), if we relax the assumption of independence between the cluster effects and the regressors;

$E[\lambda_c / X_c] \neq E[\lambda_c]$, this is a fixed effects model.

More importantly, Hausman test of null of independence between the cluster (or stratification) effects and regressors can be used to select between the random and fixed effects models (Wooldridge, 2001).

5.1. NIGERIAN DEMOGRAPHIC AND HEALTH SURVEY (NDHS)

The Nigerian Demographic and Health Survey (NDHS) provide estimates of national health and family planning statistics. The survey was designed to provide estimates for Nigeria as a whole, for urban and rural areas and for the six geopolitical regions. The survey is always conducted on yearly basis. The 2005 NDHS data has been used in this work. The health variable used as the dependent variable is the nutritional status of children and some selected variables as the socio-economic status indicators (regressors). A child's nutritional status is assessed by comparing height and weight measurements against an international standard. By this standard many children are malnourished (NDHS, 2005). Three categories of nutritional status are identified, and they are; stunting (height-for-age), wasting (weight-for-height) and underweight (weight-for-age). The socio-economic variables used are wealth index, ideal number of children, and size of the baby at birth, current age of the child, sex of the child, the birth order, parents' index of height for age, total number of children ever born, educational level and religion of the parents.

6. STANDARD ERROR ADJUSTMENT FOR CLUSTERING AND STRATIFICATION

A multiple regression using stunting (height-for-age) as the nutritional status on the selected independent variable using ordinary least square (OLS) method is obtained. In

MULTIVARIATE REGRESSION IN COMPLEX SURVEY DESIGN

addition to OLS estimates, we present analysis that gives standard errors with various degree of adjustment for stratification, cluster and both combined in the model. The stratification variable used in the model is the region, which constitute the six geopolitical zones in the country. The clustering variable is the household unit. STATA package is used to analyze the data because of its uniqueness in handling complex survey design.

The OLS point's estimates of the independent variables (coefficients) and the standard errors are given in column 1 and 2 respectively in Table 1. Relative to simple OLS standard errors, the standard errors adjusted for cluster, stratification, and both stratification and cluster combined are shown in columns 3, 4, and 5 respectively. The overall model summary (R-Square) and sample size are also indicated.

	Standard errors				
	Coefficient	Unadjusted	Cluster	Stratifying	Stratifying & Cluster
Wealth index	0.1647***	0.0251	0.0289	0.0248	0.0289
Ideal no of children	-0.0373**	0.0113	0.0134	0.0116	0.0134
Child size at birth	-0.1279***	0.0440	0.0462	0.0438	0.0463
Current age of child	-0.3514***	0.0277	0.0315	0.0286	0.0367
Sex	-0.2293***	0.0599	0.0600	0.0600	0.0600
Birth order	-0.4960***	0.0660	0.0685	0.0680	0.0674
Parents index					
Of height for age	0.1909***	0.0296	0.0321	0.0304	0.0319
Total no of children	0.5422***	0.0664	0.0702	0.0680	0.0689
Educational level	0.1055***	0.0254	0.0264	0.0240	0.0263
Religion	-0.0664***	0.0664	0.0753	0.0654	0.0730
Constant	-0.4334**	0.1946	0.2149	0.1924	0.2149
Sample size = 5,138	R ² = 0.1559	R ² = 0.1596	R ² = 0.1596	R ² = 0.1596	

* sign. at 0.10, ** sign. at 0.05 and *** sign. at 0.01

Table 1. Regression analysis of stunting on some socioeconomic status

7. RANDOM AND FIXED EFFECTS FOR STRATIFICATION AND CLUSTER

We also examined and compared effects of cluster as a fixed or random in relation to when it is only adjusted for in the ordinary least square (OLS) model. For consistency, we used the dependent and independent variables used earlier. Table 2 gives the coefficients for OLS, fixed as well as random effects model in columns 1, 3, and 5 respectively, while columns 2, 4, and 6 give their corresponding standard errors in that order. Table 3 gives the corresponding result for stratification. Both Hausman and Breuch-Pagan Lagrange Multiplier (B-P LM) tests of significance of the stratifying or clustering variable are indicated in the two tables.

Coefficients & Standard errors						
	Cluster		Random Effects		Fixed Effects	
	Coeff.	Adj. S.E (OLS)	Coeff.	S.E	Coeff.	S.E
Wealth index	0.1647***	0.0289	0.1607***	0.0295	0.0949*	0.0451
Ideal no of children	-0.0373**	0.0134	-0.0302**	0.0115	-0.0180	0.0126
Child size at birth	-0.1279***	0.0462	-0.1663***	0.0443	-0.2014***	0.0471
Current age of child	-0.3514***	0.0315	-0.3550***	0.0274	-0.3693***	0.0287
Sex	-0.2293***	0.0600	-0.2356***	0.0588	-0.2406***	0.0609
Birth order	-0.4960***	0.0685	-0.4994***	0.0651	-0.5264***	0.0679
Parents index						
Of height for age	0.1909***	0.0321	0.1847***	0.0299	0.1792***	0.0322
Total no of children	0.5422***	0.0702	0.5393***	0.0655	0.5555***	0.0683
Educational level	0.1055***	0.0264	0.1034***	0.0263	0.0770***	0.0296
Religion	-0.0664***	0.0753	-0.3984***	0.0737	-0.1564	0.0998
Constant	-0.4334**	0.2149	0.4403*	0.2127	-0.5542*	0.2679
Sample size = 5,138						
R ² = 0.1596		R ² = 0.1589		R ² = 0.1434		
				B-PLM = 81.27		p-value = 0.0000
				Hausman = 33.02		p-value = 0.0003

* sign. at 0.10, ** sign. at 0.05 and *** sign. at 0.01

Table 2. Regression analysis of stunting on some socio-economic status for cluster

MULTIVARIATE REGRESSION IN COMPLEX SURVEY DESIGN

Coefficients & Standard errors						
	Stratification		Random Effects		Fixed Effects	
	Coeff.	Adj. S.E (OLS)	Coeff.	S.E	Coeff.	S.E
Wealth index	0.1647***	0.0248	0.1647***	0.0251	0.1745***	0.0250
Ideal no of children	-0.0373**	0.0116	-0.0373**	0.0113	-0.0179	0.0114
Child size at birth	-0.1279***	0.0438	-0.1279**	0.0440	-0.2014***	0.0438
Current age of child	-0.3514***	0.0286	-0.3514***	0.0277	-0.3484***	0.0271
Sex	-0.2293***	0.0600	-0.2293***	0.0599	-0.2245***	0.0587
Birth order	-0.4960***	0.0680	-0.4960***	0.0660	-0.4911***	0.0647
Parents index Of height for age	0.1909***	0.0304	0.1909***	0.0296	0.1796***	0.0291
Total no of children	0.5422***	0.0680	0.5422***	0.0664	0.5264***	0.0651
Educational level	0.1055***	0.0240	0.1055***	0.0254	0.0837***	0.0252
Religion	-0.0664***	0.0654	-0.4553***	0.0664	-0.1892	0.0740
Constant	-0.4334**	0.1924	-0.4334*	0.1946	-0.5542***	0.1974
Sample size = 5,138 $R^2 = 0.1596$ $R^2 = 0.1596$ $R^2 = 0.1493$ B-PLM = 81.27 p-value = 0.0000 Hausman = 33.02 p-value = 0.0003						

* sign. at 0.10, ** sign. at 0.05 and *** sign. at 0.01

Table 3. Regression analysis of stunting on some socio-economic status for stratification

8. DISCUSSION OF RESULTS

Adjusting for stratification alone (Table 1 column 4) tends to inflate the standard errors in some cases when compared with conventional OLS standard errors. In some cases, the adjustment is slightly downward. In no case does the adjustment change the level of significant. Adjusting for cluster and not for stratification (Table 1 Column 3) has greater impact than stratification. In all the cases, the adjustment is upward and it actually changed the level of significant of constant term from being significant to not significant. Meanwhile, adjusting for both stratification and cluster (Table 1 column 5) shows that there is no difference between adjusting for cluster and for both stratification and cluster. The implication is that given the adjustment for clustering, the marginal impact of stratification adjustment is small (columns 3 and 5 of Table 1).

Tables 2 and 3 give the results of examining the fixed and random effects models for clustering and stratification variables respectively. For clustering variable, apart from having different coefficients for both fixed and random effects models compared with OLS, fixed effect model indicated two variables (ideal number of child & religion) to be insignificant whereas all the variables are significant in both OLS and random effects models. Similar result was obtained for stratification variable except that only ideal number of children is reported to be insignificant for the fixed effect model.

CONCLUSIONS

In multivariate regression analysis of health survey, especially in a complex survey design, there is need for adjustment of standard errors of estimates of the regressors coefficient. Adjusting for clustering is more important, since there is no significant difference between adjusting for clustering and adjusting for both stratification and clustering combined. Though the adjustment may not and in most cases do not change significant of regressors.

Also, determining whether the clustering or classification variable is exogenous or endogenous is very important. A fixed effect model as seen from the analysis is significantly different from random effect model (Hausman Fixed test). In both cases (Clustering and Stratification), the random effects model indicated the same variables reported to be significant by the OLS while the fixed effects model is slightly differed in the variables selected.

REFERENCES

1. Cochran, W. G. Sampling Techniques. New York, John Wiley and Sons. 3rd Edition, 1977. (ISBN:0-047-16240-X, QA: 276.6.c6, 1977, 001.4.222, 77-728)
2. Ferguson, J. A. and Corey, P. N. Adjusting for Clustering in Survey Design; The Annals of Pharmacotherapy; Vol. 24(3), 1990, pp. 310-313.

MULTIVARIATE REGRESSION IN COMPLEX SURVEY DESIGN

3. Ke, H., Chang, L., Carsten, R. and Wing, H. W. Detect and Adjust for Population Stratification in Population-Based Association Study Using Genomic Control Markers. An Application of Affimatrix Genechip Human Mapping 10k Array; European Journal of Human Genetics; Vol. 12, 2004, pp. 1001 – 1006.
4. Nigerian Demographic and Health Survey (NDHS). National Population Commission [Nigeria] and ORC Macro. 11785 Beltsville Drive, Calverton, MD 20705 USA, 2005.
5. Okafor, C. F. Sample Survey Theory with Applications; Afro-Orbis Publications Ltd, University of Nigeria, Nsukka, Nigeria. 1st Edition, 2002. (ISBN: 978-35253-7-9)
6. Owen, O. D., Eddy, V. D., Adam, W. and Magnus, L. Analysing Health Equity Using Household Survey data: A guide to Techniques and their implementation; World Bank Publication; WBI Learning Resources Series, 2007. (ISBN: 978-0-8213-6933-3. eISBN: 978-0-8213-6934-0. DOI: 10.1596/978-0-8213-6933-3)
7. Raj, D. and Chandhok, P. Sample Survey Theory; Nawsa Pub. House, London. 1st Edition, 1999. (ISBN: 81-7319-136-0)
8. Wooldridge, J. M. On the Robustness of Fixed Effects and Related Estimators in Correlated Random Coefficients Panel Data Models, 2001; <http://econmes.jamu.edu/workshops/economics>
9. Wooldridge, J. M. Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity; Journal of Applied Economics; Vol. 20(1), 2005, pp. 39 – 54.