International Journal of u- and e- Service, Science and Technology
Vol.8, No. 10 (2015), pp.65-74
http://dx.doi.org/10.14257/ijunesst.2015.8.10.07

# Analysis of Regression Based on Sampling Weights in Complex Sample Survey: Data from the Korea Youth Risk Behavior Web-Based Survey

Haewon Byeon[1], Heekyung Jin[2], Seonghun Yu[3] and Sunghyoun Cho[4*]

[1] *Dept. of Speech Language Pathology & Audiology, Nambu University, Gwangju, South Korea, byeon@nambu.ac.kr*
[2] *Dept. of Physical Therapy, Nambu University, Gwangju, South Korea vksclfl@nate.com*
[2] *Dept. of Physical Therapy, Gwng-ju Trauma Center, Gwangju, South Korea, yshjj18@hanmail.net*
[4*] *Dept. of Physical Therapy, Nambu University, Gwangju, South Korea geriatricpt1@naver.com*

***Abstract***

*In numerous epidemiological data, complex sample survey including stratification variables and cluster variables is used as a sampling method. This study compared the differences in mean and significant probability depending on the application method of weights when a chi-square test is performed using Complex Sample Survey Data. Multiple regression and complex sample regression analysis were used as analyzing methods. Youth Risk Behavior Web-based Survey (KYRBS) 2012 was used as data source. As a result of multiple regression and complex sample regression analysis, there were differences in mean and significant probability. The result of this study confirmed that in the analysis of data from complex sample surveys, biased variance (standard error) may be drawn out when using Simple Random Sampling Design. When using complex sampling survey data, in order to minimize the error for samples and to have study results that represent overall population, Complex Sampling Design is required.*

***Keywords:*** *Complex Sample Survey, complex sample regression analysis, cluster sampling*

## 1. Introduction

A disease is a complex body of various risk factors. Therefore, for an effective prevention, it is important to clearly elucidate and manage potential risk factors related to the disease [1]. With the rise of interest in prevention of diseases and promotion of public health, studies using epidemiological data have drastically increased. For instance, in 2011, a number of studies using KYRBS, which is a typical epidemiological survey in Korea, was only 13 while in 2014, the number increased twofold to 21 [2]. As the samples surveyed in these studies were extracted from local population, they are more reliable than other studies using a local region or hospital(s) in investigating the risk factors of diseases.

However, since a large number of domestic studies using national epidemiological data have been conducted in unweighted single sample analysis, there are possibilities of drawing out biased results [3]. In addition, even the studies applying weights may have errors in results [5] if they conduct analysis in the frequency weight method of existing single sample analysis, because the data source and national statistical survey, drew out samples by complex sample design using stratification and cluster sampling [4]. Although preceding studies on complex

ISSN: 2005-4246 IJUNESST

sample design [5-9] have raised the possibility of error due to application method of weights, they have been limited to theoretical studies and there have been few studies that verified the difference in results depending on analysis method.

This study compared unweighted simple random sampling analysis, weighted single sample analysis, and complex sample analysis which applied stratified weights by using the obesity data of the national statistical survey and analyzed if there is statistical difference in drawn-out potential risk factors.

## 2. Methods

### 2.1 Data and Participants

This study used 2012 Youth Risk Behavior Web-based Survey (KYRBS). KYRBS was jointly conducted by Ministry of Education, Science and Technology, Ministry of Health and Welfare and Korea Center for Disease Control and Prevention for the purpose of preparing countermeasures for youth's health risk behaviors [10]. Raw data were used after the authorization of Korea Center for Disease Control and Prevention. The population for 2012 KYRBS was students in middle and high school students across the country as of April 2011. The objects of the survey were 76,980 students of 400 middle schools and 400 high schools, and participants were 74,186 students in 797 schools.

The sample extraction process can be divided into stages of population stratification, samples distribution, and sample extraction. In the population stratification stage, in order to minimize sampling error, the population was divided into 135 strata by using region and type of schools (middle school, academic high school, technical high school) as stratification variables. In the sample distribution stage, distribution was made by applying proportional allocation so that the population composition by stratification variable corresponds with sample composition. Aside from that, two-stage cluster sampling method was used and the primary sampling unit was school and secondary sampling unit was class. For primary sampling, after the school lists in the population were aligned by strata and sampling intervals were calculated, sample schools were selected according to the systematic sampling method. For secondary extraction, one class per grade was randomly sampled from selected sample schools. Whole students in the selected sample class were surveyed while students with long absences, students with special needs, and words decoding disorder were excluded from sample students.

Objects of the survey were 15 areas including smoking, drinking, and obesity. The method of the survey was assigning a computer per each student in the computer room of the schools connected to the Internet and having them respond to self-administered questionnaires on randomly assigned seats.

Objects of analysis in this study were 36,889 high school students at the age of 15 through 19 who completed 2012 KYRBS.

### 2.2 Measures

Outcome was defined as Body Mass Index (BMI), which is a continuous variable. Formula to calculate BMI is as below:

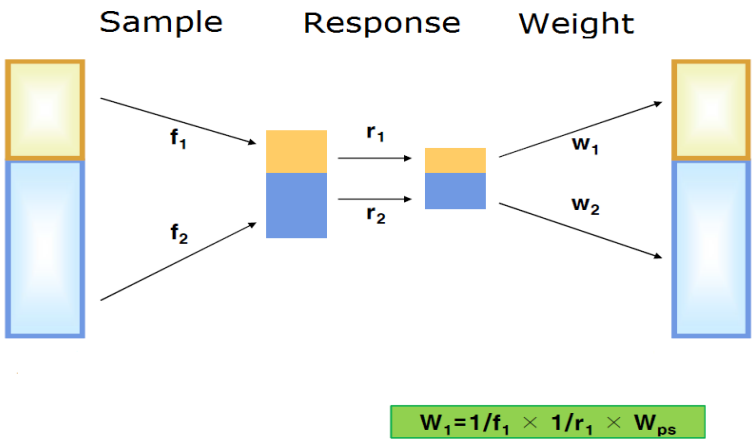
$$\mathrm{BMI} = \frac{\mathrm{mass(kg)}}{(\mathrm{height(m)})^2}$$

Explanatory variables included types of school (vocational school, academic school), grade (1st, 2nd, 3rd), subjective recognition of body type (obese, average, underweight), subjective health (healthy, average, unhealthy), economic level (high,

intermediate, low), city of residence (metropolitan, mid- to small-sized city, rural area) and academic scores (high, intermediate, low). All explanatory variables were surveyed in categorical type.

### 2.3 Sampling Weight




Potential risk factors for obesity were analyzed by using unweighted simple random sample analysis, single sample analysis applying frequency weight, and complex sample analysis applying stratified secondary weight respectively. Weights of KYRBS were applied so that subjects of the survey represent overall middle and high school students in Korea [93]. Summary of weight is as follows:

Weighted value equals inversion of extraction rate multiplied by inversion of response rate multiplied by ex post correction rate.

*Weighted value = (1/extraction rate) * (1/response rate) * ex post correction rate*

Extraction rate is calculated reflecting the sample extraction process of sample design and equals extraction rate of sample schools multiplied by extraction rate of sample classes. Extraction rate is calculated with inversion of extraction rate so that it can represent population.

*1/extraction rate = number of population schools/number of sample schools * number of classes per grade in sample schools*

For response rate, response rate per grade in sample schools was used and was calculated in rate of subjects participating in the survey among number of subjects (number of students on the roll book at the date of survey) per grade in sample schools.

Ex post correction rate of weights was calculated in such a way that sum of weights by gender, type of schools (middle school, academic high school, vocational school) and grade equals the number of students in middle and high schools in Korea as of April 2009.

*Ex post correction rate of weights = number of students in regional group by gender, type of schools and grade in population / sum of weights by gender, type of schools and grade in regional group*

### 2.3 Statistical Analysis

Simple random sampling analysis was conducted by using surveyed sample as it is without applying weights. Function to estimate mean and variance of simple random sampling analysis is as below:

$$\hat{\bar{Y}} = \frac{\sum_{k=1}^{n} y_k}{n} \qquad \hat{V}(\hat{\bar{Y}}) = \frac{\sum_{k=1}^{n} (y_k - \hat{\bar{y}})^2}{n-1}$$

Single sample analysis was conducted by applying frequency weights.

Considering the sample design of KYRBS, complex sample analysis was conducted by applying all of cluster variables, stratified variables and weights. Function for mean and variance of complex sample analysis is as follows:

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi.} - \overline{e_{h..}})^2$$

$$f_h = n_h / N_h$$

$$e_{hi.} = [\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij} - \hat{\bar{Y}})]^2 / w_{..}$$

$$\overline{e_{h.}} = (\sum_{i=1}^{n_h} e_{hi.}) / n_h$$

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} \times y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}}$$

Potential risk factors were verified by multiple regression in a simple random sample analysis and a single sample analysis, while they were verified by complex sample regression analysis in complex sample analysis.

Function for regression model of single sample analysis is as below:

$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ , } i=1, \ldots, n$$

$$\beta_0, \ \beta_1 \text{ s.t minizing the sum of squared errors } \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_2 x_i)]^2$$

Function for regression model of complex sample analysis is as below:

$$\sum_{i=1}^{N} [y_i - (B_0 + B_1 x_i)]^2 \text{ , } i=1, \ldots$$

All analysis used MINITAB version 16 (Minitab Inc., State College, Pennsylvania, USA).

## 3. Results

### 3.1 Characteristics of Population based on Weights

Characteristics of population based on weights are presented in Table 1. As a result of technical analysis, there were differences in rates between unweighted simple random sample analysis and weighted complex sample analysis in all variables. More importantly, the rate of metropolitan city was higher in simple random sample analysis whereas the rate of small and mid-sized city was higher in weighted complex sample analysis.

As in the result of complex sample analysis, the rates of academic high school (77.0%), small and mid-sized city (48.3%), low academic grade (39.5%), average subjective happiness (53.1%), good subjective health (64.8%), subjective obesity (39.1%), intermediate economic level (47.3%), normal BMI (59.4%) were high in the population.

### 3.2 Prevalence Rate of Obesity

The prevalence rates of obesity depending on whether weight and its characteristics were applied are presented in Table 2. There was difference in the prevalence rate between simple random sample analysis and weighted complex sample analysis. As the result of complex sample analysis, the prevalence rate of obesity was high in 3rd grade (25.8%), vocational schools (25.0%), metropolitan city (23.3%), low-grade group (24.2%), group subjectively recognizing themselves as unhappy (24.1%), subjective obesity (53.7%), low economic level (25.0%).

### 3.3 Comparison of Significant Probability between Single Sample Analysis and Complex Sample Analysis

Comparison of significant probability between single sample analysis and complex sample analysis to explore risk factors of obesity is presented in Table 3. As in the result of regression analysis, there was a difference in drawn-out significant probabilities from unweighted simple random sample analysis, single sample analysis to which only frequency weight was applied, and complex sample analysis. In the case of simple random sample analysis, potential risk factors for obesity were grade, type of school, subjective health, subjective body weight, and economic level ($P<0.05$). On the other hand, in single sample analysis with only frequency weight applied, all the variables were potential risk factors for obesity ($p<0.05$). In complex sample analysis, potential risk factors for obesity were grade, academic score, subjective happiness, subjective health, subjective body weight, and economic level ($p<0.05$). While academic score and subjective happiness were not significant risk factors for obesity in simple random sample analysis, they were significant risk factors in complex sample analysis. On the contrary, type of school was a significant risk factor for obesity in simple random sample analysis whereas it was not a significant risk factor in complex sample analysis.

**Table 1. General Characteristics of Korean High School Student based on sampling weight**

| Characteristics | Unweighted % | Weighted % |
|---|---|---|
| Grade | | |
| 1st | 33.8 | 33.5 |
| 2nd | 33.4 | 33.6 |
| 3rd | 32.9 | 32.9 |
| Types of school | | |
| Academic school | 77.1 | 77.0 |
| Vocational school | 22.9 | 23.0 |
| Residing city | | |
| Metropolitan | 46.1 | 45.2 |
| Mid- to small-sized city | 43.2 | 48.3 |
| Rural area | 10.7 | 6.6 |
| Academic scores | | |
| High | 31.8 | 31.8 |
| Intermediate | 28.6 | 28.7 |
| Low | 39.6 | 39.5 |
| Subjective happy | | |
| Happy | 33.1 | 33.2 |
| Average | 53.0 | 53.1 |
| Unhappy | 13.9 | 13.8 |
| Subjective health | | |
| Healthy | 64.5 | 64.8 |
| Average | 26.8 | 26.6 |
| Unhealthy | 8.8 | 8.6 |
| Subjective recognition on body type | | |
| Average | 32.6 | 32.6 |
| Underweight | 28.2 | 28.3 |
| Obese | 39.2 | 39.1 |
| Economic level | | |
| High | 24.6 | 25.0 |
| Intermediate | 47.7 | 47.3 |
| Low | 27.8 | 27.6 |
| BMI | | |
| Average | 59.5 | 59.4 |
| Underweight | 17.6 | 17.5 |
| Obese | 23.0 | 23.1 |

**Table 2. Prevalence of Obesity**

| Variables | Unweighted % | Weighted % |
|---|---|---|
| Grade | | |
| 1st | 21.2 | 21.0 |
| 2nd | 22.3 | 22.6 |
| 3rd | 25.5 | 25.8 |
| Types of school | | |
| Academic school | 22.4 | 22.6 |
| Vocational school | 24.8 | 25.0 |
| Residing city | | |
| Metropolitan | 23.1 | 23.3 |
| Mid- to small-sized city | 22.7 | 22.9 |
| Rural area | 23.2 | 23.1 |
| Academic scores | | |
| High | 21.8 | 22.2 |
| Intermediate | 22.5 | 22.7 |
| Low | 24.2 | 24.2 |
| Subjective happy | | |
| Happy | 22.8 | 23.2 |
| Average | 22.9 | 23.1 |
| Unhappy | 23.6 | 23.0 |
| Subjective health | | |
| Healthy | 22.6 | 22.7 |
| Average | 23.6 | 23.7 |
| Unhealthy | 23.4 | 24.1 |
| Subjective recognition on body type | | |
| Average | 7.3 | 7.5 |
| Underweight | 0.4 | 0.4 |
| Obese | 53.4 | 53.7 |
| Economic level | | |
| High | 24.0 | 24.5 |
| Intermediate | 21.3 | 21.3 |
| Low | 24.8 | 25.0 |

**Table 3. Comparison of Significant Probability between Single Sample Analysis and Complex Sample Analysis to Explore Risk Factors of Obesity**

| Variables | Unweighted | Frequency weighted | Complex weighted |
|---|---|---|---|
| Grade | <0.001 | <0.001 | <0.001 |
| Types of school | 0.017 | <0.001 | 0.237 |
| Residing city | 0.563 | <0.001 | 0.637 |
| Academic scores | 0.413 | <0.001 | 0.012 |
| Subjective happy | 0.087 | <0.001 | <0.001 |
| Subjective health | 0.005 | <0.001 | <0.001 |
| Subjective recognition on Body type | <0.001 | <0.001 | <0.001 |
| Economic level | <0.001 | <0.001 | <0.001 |

## 4. Discussion

Samples of KYRBS in this study had differences in drawn-out characteristics of population and prevalence rate depending on the application of weights. Sampling of KYRBS was based on a 3-stage stratified sampling method of primary sampling unit (Dong, Eup, Myeon), secondary sampling unit (survey area), and tertiary sampling unit (household) [11]. This stratified sampling method is identically used in most national statistical surveys such as Community Health Survey, Working Condition Survey and Korean Longitudinal Study of Ageing (KLoSA), as well as Health Care Survey [12-14]. When estimating population in national data extracted by stratified sampling, serious bias may arise if analysis is conducted while ignoring weights [6]. In this study, there was also a difference in population rate and tendency in the prevalence rate of obesity. Lee [4] pointed out that, in the case of data drawn out by complex sample, serious bias may occur and variance of estimate amount may be underestimated if sample design such as stratification, cluster sampling and weights is not reflected in analyzing stage. Therefore, when using national statistical data surveyed by stratified sampling method in a study, only the application of complex sampling method enables expansion of drawn-out results to overall population, which thus in turn enables reliable interpretation of the results. In particular, in order to estimate unbiased prevalence rate of a disease, a weighted method is a necessity in the analysis process.

As in the result of analysis on potential risk factors for obesity by using simple random sampling analysis, single sampling analysis applying frequency weight, and complex sampling analysis, in the case of single sampling analysis with only frequency weight applied, all the variables were overly predicted to be significant risk factors while, in the case of simple random sampling analysis with no weights applied and complex sampling analysis, there were differences in significance level and results. Traditionally, a linear regression model has been performed to analyze potential risk factors in continuous data while cluster sampling does not follow t-distribution since correlation exists among observed values [8, 9]. In addition, Lv, *et al.*, [7] also pointed out that when only frequency weight is applied to cluster sampling data, there is a possibility that significance levels of all variables may become smaller as the size of the sample becomes bigger [6]. Nevertheless, many studies do not gain accurate analysis results as they perform statistical analysis, wrongly assuming the data from stratified sampling as those from simple random sampling [5, 8, 15]. As an alternative to this problem, Lee [16] suggested using the Rao-scott chi-square test as a categorical analysis method for complex sampling data and, by using data from Third National Health and Nutrition Examination Study of the U.S., proved that Rao-scott chi-square test can present more stable statistical figures than the Wald test. To sum it up, in order to draw out unbiased estimates from national data, it is necessary for future studies to use the Rao-scott chi-square test or regression analysis of complex sample survey data.

## 5. Conclusion

When using national statistical data surveyed by using stratified sampling method, tests utilizing complex sample analysis should be used to have the results of the study represent the population at large. Aside from that, in the case when only frequency weight applied, extra attention is required as there is a high possibility of over-interpretation of the results.

## References

[1] C. Tudor-Locke, “Walk more (frequently, farther, faster): The perfect preventive medicine”, Preventive Medicine, vol. 55, **(2012)**, pp. 540-541.
[2] K. W. Oh, “Introduction to database for public health research”, Epidemiology & Biostatistics Training

Program, vol. 86, **(2011)**.
[3] J. W. Sakshaug and B. T. West, "Important considerations when analyzing health survey data collected using a complex sample design", American Journal of Public Health, vol. 104, **(2014)**, pp. 15-16.
[4] K. J. Lee, "Comparison of regression model approaches fitted to complex survey data", Survey Research, vol. 2, **(2000)**, pp. 73-86.
[5] H. Byeon, "Comparative analysis of unweighted sample design and complex sample design related to the exploration of potential risk factors of dysphonia", Journal of the Korea Academia-Industrial Cooperation Society, vol. 13, **(2012)**, pp. 2251-2258.
[6] W. Kalsbeek and G. Heiss, "Building bridges between populations and samples in epidemiological studies", Annu Rev Public Health, vol. 21, **(2000)**, pp. 147-169.
[7] J. Lv, P. P. He, W. X. Tu and L. M. Li, "Estimation of sampling error on data from cluster sample survey", Zhonghua Liu Xing Bing Xue Za Zhi, vol. 29, **(2008)**, pp. 78-80.
[8] K. Nam and B. Cho, "$\chi^2$ tests for categorical data in complex sampling surveys", The Korean Economic Review, vol. 20, **(1993)**, pp. 277-288.
[9] J. N. K. Rao and A. J. Scott, "The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way Tables", Journal of the American Statistical Association, pp. 76, **(1981)**, pp. 221-230.
[10] "Ministry of Health and Welfare", Korea Youth Risk Behavior Web-based Survey 2012, Seoul: Ministry of Health and Welfare, **(2013)**.
[11] K. O. Cho, "Physical activity and suicide attempt of South Korean adolescents-evidence from the eight Korea youth risk behaviors web-based survey", Journal of Sports Science & Medicine, vol. 13, **(2014)**, pp. 888-893.
[12] "Korea Centers for Disease Control and Prevention", "2009 Community Health Survey, Seoul: Korea Centers for Disease Control and Prevention", **(2012)**.
[13] "Ministry of Employment and Labor", "2005 Survey report on labor conditions at small size extablishments", Gwacheon: Ministry of Employment and Labor, **(2006)**.
[14] Korea Labor Institute, "2008 Korean Longitudinal Study of Ageing", Seoul: Korea Labor Institute, **(2009)**.
[15] H. Byeon, H. Jin, S. Yu and S. Cho, "Regression analysis of complex sample survey data: results from a national survey", Advanced Science and Technology Letters, vol. 104, **(2015)**, pp. 93-96.
[16] S. Lee, "A study on varian stability of quadratic form test statistics under a complex survey design", Journal of The Korean Official Statistics, **(1999)**, pp. 123-138.

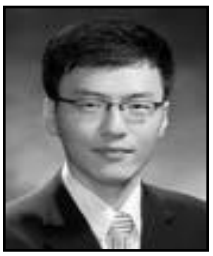
## Authors

**Haewon Byeon**, received his DrSc degree in Biomedical Science from Ajou University School of Medicine. He is a professor in Department of Speech Language Pathology & Audiology and director of Speech Language Pathology Center in Nambu University. His recent interests focus on health promotion and biostatistics.

**Heekyung Jin**, received her PhD degree in Physical Therapy from Seonam University. Her recent interests focus on biomechanics and electrophysiology.

**Sunghun Yu**, received his PhD degree in Physical Therapy from Dongshin University. His recent interests focus on Psychology and manual therapy



**Sunghyoun Cho**, received his PhD degree in Physical Therapy from Daegu University. He is a professor in Department of Physical Therapy in Nambu University, Gwangju, Republic of Korea. His recent interests focus on Biomechanics and Therapeutic exercise.