The Korean Communications
in Statistics Vol. 9, No. 3, 2002
pp. 775-786

Design and Weighting Effects in Small Firm Survey in Korea[1)]

Keejae Lee[2)], James M. Lepkowski[3)]

Abstract

In this paper, we conducted an empirical study to investigate the design and weighting effects on descriptive and analytic statistics. The design and weighting effects were calculated for estimates produced from the 1998 small firm survey data. We considered the design and weighting effects on coefficients estimates of regression model using the design-based approach and the GEE approach.

Keywords : Design effect, Weighting effect, Survey regression model, GEE approach

1. Introduction

Most of the national large surveys use a complex design with stratification, clustering and unequal weights. This is mainly due to the large costs involved in simple random sampling. The effect of complex sample design on an estimator can be measured by the design effect, which is the ratio of the variance of the estimator under the complex sample design to the variance calculated as if the sample data came from simple random sampling. Kish and Frankel(1974) presented some empirical results evaluating design effects for estimators such as means, proportions and linear regression coefficients. Design effects for complex sample design can be used in various ways. For example, they are used to determine the effective sample size of complex sample design. Design effect can also be used to approximate the sampling variances of statistics from complex surveys when one does not calculate the variance of estimator in usual method.

In many of the national establishment surveys, it is also important to make estimates for

1) This study was carried out when the first author was visiting the University of Michigan in 2000 under the financial support from Korea Science and Engineering Foundation.

2) Associate Professor, Department of Information Statistics, Korea National Open University, 169 Dongsung-Dong, Jongno-Gu, Seoul, 110-791, Korea.
E-mail: kjlee@mail.knou.ac.kr

3) Senior Research Scientist, Institute for Social Research, The University of Michigan, Ann Arbor MI 48106-1248
E-mail: jimlep@umich.edu

the domain such as industry classification, occupations, gender etc. This requires differential sampling rates so as to obtain adequate sample sizes for various domains. Differential sampling rates require weighting of the sample data. Weighting may also be introduced to compensate for differential non-response. Ignoring the sample weights in an analysis can lead to substantial bias.

In this paper, we present the design and weighting effects on several descriptive and analytic statistics in small firm survey in Korea. The small firm survey has been conducted by the Ministry of Labor since 1995. Each year in November, the data are collected from small business firms with 1-4 employees to estimate the monthly wage and hours(regular, overtime) worked in different individual characteristics of employees such as occupations, genders, educational attainment etc. This survey has a stratified one-stage cluster sampling design. The sample is composed of 14,942 firms(clusters) with a total of 33,116 employees.

In section 2, we discuss the design and weighting effects on descriptive statistics and conduct an empirical study to measure the inefficiency due to weighting. To compute the standard errors in this study, we use the Taylor linearization method applied in the SUDAAN. In section 3, the wage is modeled using both design-based approach and the Generalized Estimating Equation(GEE) approach as a linear function of industry classification, occupation, gender, educational attainment, etc. In section 4, we consider design and weighting effects on estimation of linear regression coefficients using both the design-based approach and the GEE approach. Section 5 gives the concluding remarks.

2. Design and Weighting Effects on Descriptive Statistics

The sampling frame of the small firm survey is the Business Register of the Ministry of Labor, except for the firms which belong to agriculture, forestry, fishing and hunting industry. The population of the survey consists of 845,376 firms, which are stratified into 52 categories by the Korean Standard Industry Classification.

The sampling fractions vary from stratum to stratum. Each employee in the sample receives a weight which represents the respondent's contribution to the entire population and is used to derive unbiased estimates for characteristics of interest. The weight is derived as the product of three factors: a design weight, a non-response adjustment and poststratification adjustment. The poststratification adjustment cells in this survey are defined by the sampling strata cross-classified by gender. The number of the poststratification adjustment cells is 104. The descriptive statistics on sample weights are given in Table 1.

Table 1. Descriptive statistics on sample weights

Percentile	Male (n=19,580)	Female (n=13,536)	Total (n=33,116)
Minimum	2.78	1.00	1.00
25%	8.79	10.76	8.79
Median	19.60	40.42	26.28
75%	75.12	111.47	89.01
Maximum	154.13	123.23	154.13
Mean	41.40	56.52	41.58
CV(%)	98.95	81.34	91.89

The sample is composed of 14,942 firms with a total of 33,116 employees and the mean of employees in the sampled firm is 2.2. Considering that the wage and hours worked are similar in the same firm, we may expect that intra-cluster correlations can be high in most study variables.

Table 2 displays the weighted means, design effects of the weighted means, unweighted means, biases of the unweighted means for the some selected study variables. The design effect for the regular hours worked variable is largest as the hours worked variables tend to be similar in the same firm in Korea. The relative bias of the unweighted mean is substantial for the overtime hours worked variable. The descriptive analysis of monthly wage can be found in Table A-1 of Appendix.

Table 2. Design effects of weighted mean and the bias of unweighted mean

Variable	Weighted mean	Deff	Unweighted mean	Rel. bias(%)
Monthly wage	880.15	2.75	905.32	2.86
ln[monthly wage]	6.70	2.86	6.73	0.45
Regular hrs worked	212.08	3.78	208.22	-1.82
Overtime hrs worked	21.04	3.57	16.49	-21.63
Age	33.69	2.38	34.57	2.61

Note. $Relative\ Bias = \frac{\bar{y}_{UW} - \bar{y}_{W}}{\bar{y}_{W}} \times 100$

The cost of using the weights in analysis is to inflate sampling error of the estimators. The methods to measure the increased variances resulted from unequal weights has been proposed by Kish(1965, 1992) and Korn and Graubard(1995). If many analyses are planned, they might

suggest differing inefficiencies due to different use of the sample weights. If this is undesirable, an approximation that is not dependent on the analysis is to calculate the inefficiency using equation (1) for sample mean.

$$Inefficiency = CV^2/(1+CV^2) = 0.4578, \quad (CV = 0.9189) \tag{1}$$

where CV is the coefficient of variation of sample weights. The approximate efficiency loss due to weighting is 45.8% in this survey.

In general, the weighted estimator with the sample weights will provide approximately unbiased estimates of the finite population quantities. Korn and Graubard(1995) proposed an formula (2) to calculate the efficiency loss in using weighted estimator instead of unweighted estimator when unweighted estimator is in fact unbiased.

$$Inefficiency = 1 - Var(\bar{y}_{UW})/Var(\bar{y}_W) \tag{2}$$

Table 3 displays the efficiency loss in using weighted mean instead of unweighted mean for some study variables. The average of inefficiencies due to weighting is 49.7%. We note that ignoring the sample weights in analysis can lead to not only substantial bias but also underestimation of sampling error of the estimates.

Table 3. Inefficiency due to weighting

Variables	Weighted		Unweighted		Inefficiency due to weighting(%)
	Mean	$s.e.(\bar{y}_W)$	Mean	$s.e.(\bar{y}_{UW})$	
Monthly wage	880.15	3.3737	905.32	2.4831	45.8
ln[monthly wage]	6.70	0.0037	6.73	0.0027	46.8
Regular hrs worked	212.08	0.3931	208.22	0.2595	56.4
Overtime hrs worked	21.04	0.3879	16.49	0.2687	52.0
Age	33.69	0.0904	34.57	0.0655	47.5

Note. The average of inefficiency due to weighting is 49.7%.

We note that it may not be reasonable to apply the Korn and Graubard's formula to assess the effect of unequal weights because of the bias of the unweighted means as shown in Table 2. To assess the inefficiency due to weighting, it is necessary to compare results between from unequal weight sample and from equal weight sample. Because there are sufficient clusters in the survey data, we can have EPSEM(Equal Probability Selection Method) subsamples of large size in the empirical study.

The selection procedure for EPSEM subsample is as follows. The idea is to resample the data from the original sample with probabilities proportional to the sample weights. The i-th cluster from the original sample is included in the subsample if a uniform (0, 1) random number is less than w_i/w_{max}, where $w_i = 1/\pi_i$ is the sample weight of the i-th cluster and w_{max} is the largest sample weight of the 14,942 sample firms. This method was used by Korn and Graubard(1999) to draw a scatterplot ignoring the sample weights. Table 4 reports the summary of the simulation results which come from 100 repetitions. For each EPSEM subsample, the unweighted sample mean is a design unbiased estimator for the population mean. The standard deviation shows the variation among unweighted means of the EPSEM subsamples.

Table 4. Inefficiency due to weighting by EPSEM subsample method

Variables	Original Sample		EPSEM Subsample		Inefficiency (%)
	$\bar{y}_w$	$Deff(\bar{y}_w)$	Mean[1] (s.d.[2])	$\overline{Deff}$[1] (s.d.[2])	
Monthly wage	880.15	2.75	869.83 (2.280)	1.50 (0.025)	45.5
ln[monthly wage]	6.70	2.86	6.69 (0.003)	1.55 (0.016)	45.8
Regular hrs worked	212.08	3.78	212.39 (0.212)	1.98 (0.016)	47.6
Overtime hrs worked	21.04	3.57	21.44 (0.245)	1.91 (0.022)	46.5
Age	33.69	2.38	33.35 (0.072)	1.29 (0.016)	45.8

Note. 1: The average of the unweighted sample means of subsamples.

2: The standard deviation of the unweighted sample means,

$\bar{m} = 6222.3$ (the average sample size of firms in the EPSEM subsamples),

$\bar{n} = 13157.6$ (the average sample size of employees in the EPSEM subsamples)

$$Inefficiency = 1 - \frac{\overline{deff}}{deff(\bar{y}_w)}$$

Table 5 reports the results by the three methods to evaluate the inefficiency due to weighting. We note that Kish's approximation formula to assess the efficiency loss due to weighting works well.

Table 5. Inefficiency summary

Variables	Kish's Method	K & G's Method	Subsampling method
Monthly wage		45.8%	45.5%
ln[monthly wage)		46.8%	45.8%
Normal working hrs	45.8%	56.4%	47.6%
Extra working hrs		52.0%	46.5%
Age		47.5%	45.8%
Average Inefficiency	45.8%	49.7%	46.2%

3. Regression Model on Wage Using the Design-based Approach and the GEE Approach

In this paper, we fit a linear regression model to data from the small firm survey using the ordinary least square method and the design-based approach respectively. We also fit the regression model accounting for exchangeable correlation structure within cluster using the GEE approach, and calculate robust standard errors.

The model-based inference is more efficient than the design-based inference when the model is correctly specified. The design-based approach is more concerned with robustness to model failure because the number of sampled elements in most surveys is fairly large. Pfeffermann and Holmes(1985) showed that the incorporation of sampling weights into estimation of regression coefficients helps to protect against the potential existence of missing regressors. Classical design-based survey methods tend to be more robust than model-based methods, but lack their efficiency. This is true not only for parameter estimates, but also for the estimation of standard errors(Korn and Graubard, 1995).

The GEE approach makes no strict distributional assumptions, but requires a specification of the mean as a linear function of predictors, and covariance, as a function of the mean and other scale parameters. With the additional requirement of a working correlation matrix that specifies the dependence of the responses, a set of generalized estimating equations are formed. Even if the specific correlation structure is misspecified, the GEE approach has been shown to yield consistent estimates of model parameters and their variances. In addition, the estimated regression coefficients are asymptotically normal.

The survey regression techniques proposed by Binder(1983) and the GEE approach introduced by Liang and Zeger(1986) are identical under the assumption of independent working correlations(Shah et al., 1997). Although the design-based approaches are valid in the presence of intra-cluster correlations, there is no attempt to use that information in computing parameter estimates. The similarities between the survey setting approach and GEE techniques

have been discussed by Bieler and Williams(1995) and Rotnitzky and Jewell(1990).

For reliable estimation of regression coefficients and their standard errors, we collapse Mining with Manufacturing and Gas, Water, Power with Construction. The result of model-based approach is obtained via SAS PROC REG and the results of both the design-based approach and the GEE approach are computed using SUDAAN. We note that industry classification, area which the firm belongs to, firm size(the number of employees) are firm level variable and the other variables are individual level.

4. Design and Weighting Effects on Regression Coefficients Estimates

In this section, we present the design and weighting effects on coefficients estimates of regression model using the design-based approach and the GEE approach.

Table A-2 displays the results of fitting the linear regression model using the model-based approach, the design-based approach and the GEE approach with exchangeable correlation structure. In this study, an exchangeable correlation structure for the employees from the same firm can be assumed. The intra-cluster correlation of the response variable(the monthly wage) within firm, ρ_Y is estimated to be 0.612. We note that the signs of regression coefficients estimates by the three approaches are the same and the patterns of estimated coefficients are similar.

In the analysis, the effects of clustering and weighting, however, are noticeable. As well known, the standard errors of coefficient estimates using the model-based approach ignoring the clustering and weighting are much smaller than the other two approaches. The standard error estimates using the design-based approach do not differ noticeably from those using the GEE approach. Specifically, the standard error estimates of the cluster level coefficients are very similar but the GEE approach provides more efficient estimation of coefficients of the individual level variables than design-based approach. Lipsitz et al. (1994) and Bieler and Williams(1995) reported similar results in clustered binary regression model.

Table 6 displays the summary of design effects in the design-based approach and the GEE approach. To conserve space, the results based on 30 regression coefficients estimates are summarized here to illustrate the design effect.

Table 6. Summary of design effects

	Design-based approach	GEE approach
Minimum	1.39	0.79
25%	2.55	2.07
Median	3.19	2.35
75%	3.53	2.86
Maximum	5.15	4.12
Mean	3.16	2.41

In addition, we note that the unweighted estimates of regression coefficients in both approaches have serious bias. Especially, the biases of unweighted estimates for the occupation variables in the design-based approach are 29.2%-91.7% and in the GEE approach, are 19.5%-50.5%.

To investigate the inefficiency due to weighting, we select 100 EPSEM subsamples from the survey data. For each EPSEM subsample, we fit the regression model using the design-based approach and the GEE approach with exchangeable correlation structure and calculate the design effect for the regression coefficient estimates. Table 7 displays the summary of the inefficiencies due to weighting in regression coefficients estimation via three methods. We note that the results by the three methods to evaluate the weighting effects are similar. The Kish's formula works reasonably in the case of regression analysis by the two approaches.

Table 7. Summary of inefficiency due to weighting in regression analysis

Approaches	Kish's Method	K & G's Method	EPSEM subsample Method
Design-based approach	45.8%	42.9%[a] 44.7%[b]	43.4%[a] 45.4%[b]
GEE approach		45.3%[a] 46.6%[b]	43.7%[a] 43.5%[b]

Note. a, b : The average and median of inefficiencies due to weighting for the 30 regression coefficients estimates

5. Concluding Remarks

In this paper, we discussed the design and weighting effects on descriptive and analytic statistics. The design and weighting effects were calculated for estimates produced from the 1998 small firm survey data. We conducted an empirical study to investigate the effect on weighting with the equal weight subsamples selected from the survey data. We considered the design and weighting effects on coefficients estimates of regression model using the design-based approach and the GEE approach.

In this study, we might gain some insights into design and weighting effects on descriptive and analytic statistics. First, the design effect has a similar pattern between descriptive and analytic statistics in the complex survey data analysis. Second, the Kish's formula approximates the inefficiency due to weighting in the cases of descriptive and analytic statistics. The Kish's simple formula works reasonably in the survey regression and GEE approach. Third, the unweighted estimates also may cause serious bias in analytic statistics.

References

[1] Bieler, G. S., and Williams, R. L. (1995). Cluster Sampling Techniques in Quantal Response Teratology and Developmental Toxicity Studies, Biometrics, 51, 764-776.

[2] Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators From Complex Surveys, International Statistical Review, 51, 279-292.

[3] Kish, L., and Frankel, M. R. (1974). Inference From Complex Samples, Journal of the Royal Statistical Society, Ser. B, 36, 1-37.

[4] Korn, E. L. and Graubard, B. I. (1995). Analysis of Large Health Surveys: Accounting for the Sampling Design, Journal of the Royal Statistical Society, Ser. A, 158, 263-295.

[5] Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*, New York: Wiley.

[6] Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1994). Performance of Generalized Estimating Equations in Practical Situations, Biometrics, 50, 270-278.

[7] Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear model, Biometriaka 73, 13-22.

[8] Pfeffermann, D., and Holmes, D. J. (1985). Robustness Considerations in the Choice of Methods of Inference for Regression Analysis of Survey Data, Journal of the Royal Statistical Society, Ser. A, 148, 268-278.

[9] Rotnitzky, A. and Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data, Biometrika, 77, 485-497.

[10] Shah, B. V., Barnwell, B. G., and Bieler, G. S. (1997). *SUDAAN User's Manual, Release 7.5.* Research Triangle Park, NC: Research Triangle Institute.

[2002년 6월 접수, 2002년 8월 채택]

Table A-1. Descriptive analysis and design effects of monthly wage

Variables	Sample size	Weighted mean	Standard error	Deff
Industry Classification				
Mining	258	934.51	30.72	1.48
Manufacturing	11250	964.12	5.15	2.60
Electricity, gas and water	241	1235.98	42.87	1.60
Construction	747	949.26	17.69	1.68
Wholesale, Retail	4050	935.40	7.44	1.55
Hotel, Restaurant	3010	796.57	7.52	1.60
Transport, communications	3600	924.73	11.66	3.25
Insurance, Finance	1976	1062.26	11.06	1.28
Real estate	3695	944.44	11.55	2.24
Education service	925	682.47	10.14	1.97
Health, Social service	692	801.44	14.19	1.79
Other service	2672	814.65	13.05	2.88
Occupation				
Senior officials, managers	1176	1358.04	35.88	3.31
Professionals	996	833.55	15.88	1.81
Technicians	3108	908.73	10.93	2.19
Clerks	10201	879.59	6.27	2.98
Sales workers	5230	812.45	5.86	1.75
Craft, related trade workers	6123	949.36	6.69	2.78
Plant and machine operators	3513	971.03	8.35	2.98
Elementary occupations	2769	753.25	9.44	3.05
Gender				
Male	19580	1034.09	5.02	3.00
Female	13536	717.02	3.27	2.59
Educational attainment				
Under middle school	5428	836.88	6.58	2.49
High school	18860	864.32	3.82	2.49
Junior college	3969	813.31	8.24	2.44
College and university	4859	1042.96	12.13	2.62
Duration of services				
< 1 year	5742	699.94	6.22	2.64
1-3 years	7345	784.36	4.67	2.25
3-4 years	3547	859.72	8.55	2.67
4-5 years	3003	886.64	7.91	1.94
5-10 years	6452	974.57	7.21	2.52
Over 10 years	7027	1126.80	9.45	3.01

Table A-2. Regression analysis via three approaches : (dep. var. : ln(monthly wage)

Variables	Model based	design-based		GEE (exchangeable)	
Industry Classification					
Manufacturing[a]	0.129[c] (0.0069[d])	0.154[c] (0.0130[d])	3.31[e]	0.156 (0.0133)	2.12
Construction[b]	0.127 (0.0110)	0.107 (0.0198)	4.37	0.121 (0.0195)	2.17
Wholesale, Retail	0.133 (0.0073)	0.147 (0.0128)	4.73	0.155 (0.0131)	2.89
Hotel, Restaurant	0.138 (0.0090)	0.122 (0.0153)	4.52	0.111 (0.0155)	2.93
Transport, Communications	0.066 (0.0076)	0.084 (0.0152)	1.63	0.086 (0.0150)	0.87
Insurance, Finance	0.221 (0.0088)	0.267 (0.0150)	1.39	0.291 (0.0154)	0.79
Real estate	0.118 (0.0077)	0.144 (0.0151)	3.13	0.139 (0.0156)	1.75
Education service	0.010 (0.0121)	0.030 (0.0181)	3.59	0.035 (0.0178)	2.19
Health, Social service	0.207 (0.0125)	0.216 (0.0175)	3.54	0.209 (0.0181)	2.20
Other service	0[f]		0[f]		0[f]
Area					
Seoul	0.073 (0.0039)	0.078 (0.0071)	3.32	0.082 0.0069)	1.72
Other metropolitan area	-0.012 (0.0040)	-0.016 (0.0073)	3.34	-0.010 (0.0073)	1.88
Rural area	0[f]		0[f]		0[f]
The firm size	0.023 (0.0016)	0.025 (0.0029)	3.45	0.024 (0.0029)	1.91
Occupation					
Senior officials, managers	0.312 (0.0111)	0.241 (0.0238)	4.09	0.275 (0.0194)	3.50
Professionals	0.169 (0.0125)	0.103 (0.0206)	3.16	0.166 (0.0197)	2.95
Technicians	0.177 (0.0091)	0.122 (0.0163)	3.21	0.156 (0.0140)	2.51
Clerks	0.143 (0.0076)	0.097 (0.0135)	2.96	0.096 (0.0114)	2.25
Sales workers	0.091 (0.0083)	0.048 (0.0139)	3.52	0.070 (0.0115)	2.40
Craft, related trade workers	0.105 (0.0080)	0.060 (0.0138)	2.52	0.084 (0.0124)	2.02
Plant and machine operators	0.130 (0.0084)	0.074 (0.0146)	2.48	0.090 (0.0126)	1.94
Elementary occupations	0[f]		0[f]		0[f]
Educational attainment					
Under middle school	-0.152 (0.0072)	-0.166 (0.0128)	3.05	-0.141 (0.0108)	2.68
High school	-0.088 (0.0052)	-0.100 (0.0092)	3.08	-0.084 (0.0077)	2.83
Junior college	-0.076 (0.0064)	-0.091 (0.0101)	2.58	-0.074 (0.0082)	2.29
College and university	0[f]		0[f]		0[f]
Gender					
Male	0.286 (0.0039)	0.261 (0.0061)	2.49	0.261 (0.0054)	2.71
Duration of services					
< 1 year	-0.242 (0.0062)	-0.228 (0.0103)	2.81	-0.260 (0.0091)	2.58
1-3 years	-0.166 (0.0056)	-0.154 (0.0093)	2.68	-0.179 (0.0081)	2.51
3-4 years	-0.125 (0.0064)	-0.115 (0.0100)	2.35	-0.131 (0.0084)	2.26
4-5 years	-0.108 (0.0067)	-0.100 (0.0103)	2.24	-0.115 (0.0088)	2.28
5-10 years	-0.065 (0.0053)	-0.067 (0.0087)	2.45	-0.073 (0.0072)	2.40
Over 10년 years	0[f]		0[f]		0[f]
Age	0.046 (0.0012)	0.043 (0.0021)	3.25	0.040 (0.0017)	3.08
Age square	-0.001 (0.0000)	-0.000 (0.0000)	3.34	-0.000 (0.0000)	3.28
Total hours of worked	0.001 (0.0000)	0.001 (0.0001)	5.05	0.001 (0.0001)	4.12
Intercept	5.308 (0.0274)	5.364 (0.0493)	3.55	5.329 (0.0429)	2.99
R^2	0.464		0.449		0.446

Note. *a* : Mining+Manufacturing, *b*: Electricity, gas and water supply+Construction,
c : Regression coefficient estimates, *d*: Standard error, *e*: Design effect,
f: Reference category