Journal of Statistical Theory and Practice (2020) 14:20
https://doi.org/10.1007/s42519-020-0086-z

ORIGINAL ARTICLE

Nonresponse Bias Adjustment in Regression Analysis

Tadayoshi Fushiki[1] · **Tadahiko Maeda**[2]

Published online: 21 February 2020

Abstract
Nonresponse is an unavoidable problem in most sample surveys. If the proportion of nonrespondents is very small, nonresponse bias may be negligible. However, nonresponse rates in sample surveys have recently increased in many countries. Thus, methods for dealing with nonresponse bias are becoming an important topic. Regression analysis is often used to analyze survey data. In this paper, we discuss regression analysis with unit nonresponse. The least square estimator of regression coefficients may be asymptotically biased if nonresponse is not ignorable. In this paper, we establish a sufficient condition that a consistent estimator of regression coefficients is obtained. This condition can be determined from a causal diagram. Furthermore, we examine the results of this study by numerical experiments.

Keywords Calibration · Causal diagram · Nonresponse bias · Regression · Sample survey · Weighted least square method

1 Introduction

1.1 Background

Nonresponse is an unavoidable problem in most sample surveys. If the proportion of nonrespondents is very small, nonresponse bias may be negligible. However, nonresponse rates in sample surveys have recently increased in many countries. Thus, estimation methods taking nonresponse into account have become more important (for example, [1, 4, 5, 11]).

Regression analysis is often used in the analysis of survey data. Linear models assumed in regression analysis are generally misspecified. The least square estimator of regression coefficients is asymptotically biased in such a situation if nonresponse

✉ Tadayoshi Fushiki
fushiki@ed.niigata-u.ac.jp

[1] Niigata University, Niigata, Japan

[2] The Institute of Statistical Mathematics/Joint Support-Center for Data Science Research, Tachikawa, Japan

is not ignorable. If a linear model is correctly specified, the least square estimator of regression coefficients is asymptotically biased when the reason for nonresponse comes from not only explanatory variables but also a response variable.

In studies of nonresponse adjustment, estimation of population totals has been focused. However, estimation of regression coefficients has received very little attention. In the present study, we establish a condition for obtaining a consistent estimator of regression coefficients by bias adjustment. We examined the results of this study by numerical experiments.

The problem of missing data appears in various forms in statistical analysis. There has been a lot of literature on this problem (for example [8]). Regression analysis with missing data has also been studied (for example [6]). The present study differs from these previous studies in that we treat unit nonresponse and use auxiliary information for bias adjustment.

1.2 Example

For simplicity, we assume sampling with replacement and an infinite population. Let Y be a response variable and X be explanatory variables. Let Z be a random variable that is set to 1 if an individual cooperates in the survey and 0 otherwise.

Here, we consider an example. We are interested in the relationship between income and age. Let Y be income and X be age. The scatter plot of (X, Y) for a sample is shown in the left panel of Fig. 1. From the figure, we can see that income increases rapidly in the twenties, but increases slowly in the fifties. Assume that the right panel of Fig. 1 is the scatter plot for respondents. The figure shows that many young individuals did not respond.

In Fig. 2, the solid line is the regression line of income on age estimated from the respondents, whereas the dotted line is that estimated from the entire sample. The slope obtained from the respondents is less than the slope obtained from the entire sample, and the intercept estimated from the respondents is larger. These results are

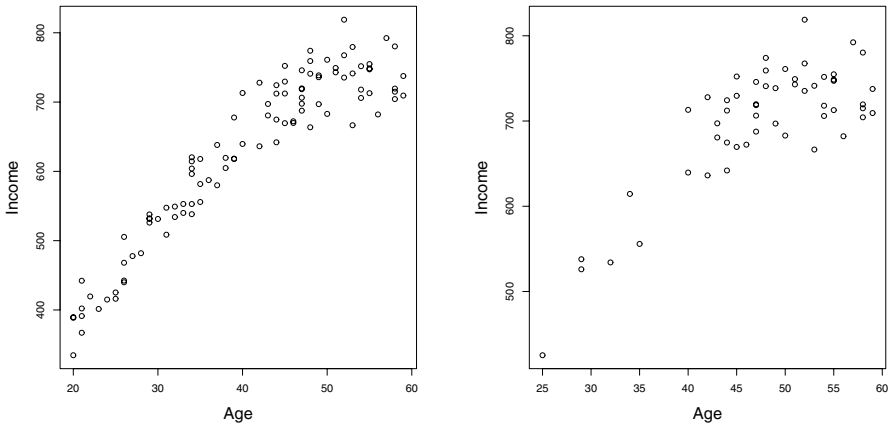

Fig. 1 Scatter plots for age and income. The left panel is for the sample and the right panel is for the respondents

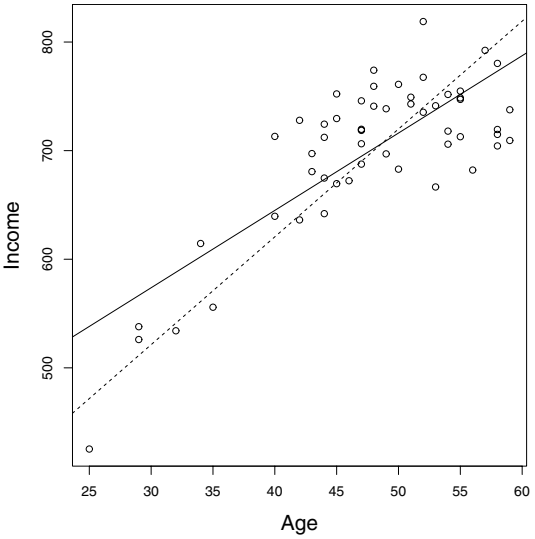

Fig. 2 The scatter plot from the right panel of Fig. 1. The solid line is estimated from the respondents and the dotted line is estimated from the entire sample

explained by the fact that the estimate from the respondents is dominated by the observations of older respondents because the response rate of old individuals is higher than that of young individuals.

In this study, we establish a condition that guarantees to obtain a bias-corrected estimator of regression coefficients.

2 Nonresponse Bias Adjustment in Regression Analysis

2.1 Condition for Obtaining a Consistent Estimator

Assume that X_1 is a part of explanatory variables X and that U is a variable set, and that the population information of (X_1, U) is available. In the following, we assume for simplicity that (X_1, U) is discrete. We denote by X_2 the remainder of X. If (Y, X_2) and Z are conditionally independent given (X_1, U), we can obtain a consistent estimator of regression coefficients, as we show below. The conditional independence is written as$(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$.

Regression coefficients for the population are given by

$$\left(E\begin{bmatrix} X_1X_1{}^T & X_1X_2{}^T \\ X_2X_1{}^T & X_2X_2{}^T \end{bmatrix}\right)^{-1} E\begin{pmatrix} X_1Y \\ X_2Y \end{pmatrix}. \tag{1}$$

Therefore, if we obtain consistent estimators for $E(X_iX_j{}^T)$ and $E(X_iY)$, a consistent estimator of regression coefficients can be obtained by applying the continuous mapping theorem (for example, [13]). By using the conditional independence, the following holds:

$$E(X_1 X_2^T) = E_{X_1,U}\{X_1 E(X_2^T | X_1, U)\} = E_{X_1,U}\{X_1 E(X_2^T | X_1, U, Z = 1)\}. \quad (2)$$

Since we can consistently estimate $E(X_2^T | X_1, U, Z = 1)$ based on data from the respondents, a consistent estimator of $E(X_1 X_2^T)$ can be obtained by using the information on the population distribution of (X_1, U). We can obtain consistent estimators of $E(X_2 X_1^T)$, $E(X_2 X_2^T)$ and $E(X_i Y)$ in the same way.

2.2 Graphical Expression for the Conditional Independence

In Sect. 2.1, it was shown that a consistent estimator of regression coefficients can be obtained if the population information of (X_1, U) is available and $(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$ holds. We now consider a graphical expression for the conditional independence.

A causal diagram is a graph expressing causal relationships between variables. We assume that causal relationships are specified and depicted by a directed acyclic graph $G = (V, E)$. Thus, the distribution of the variables of G satisfies a recursive factorization.

If (Y, X_2) and Z are d-separated by (X_1, U), then $(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$ holds. It is also known that $(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$ holds if (Y, X_2) and Z are separated by (X_1, U) in $G^{\mathrm{m}}(Y, X_2, Z, X_1, U)$, which is the moral graph of the smallest ancestral set containing $Y \cup X_2 \cup Z \cup X_1 \cup U$ (for example, [7]).

Example 1 Let Y be income and X be age. We assume that X affects Y and Z. In Fig. 3, the left panel shows the causal diagram and the right panel shows the moral graph. In the causal diagram, Y and Z are d-separated by X, thus $Y \perp\!\!\!\perp Z \mid X$ holds. Therefore, if the population information of X is available, we can obtain a consistent estimator of the regression coefficients.

Example 2 Let Y be income, X_1 be age, and X_2 be education level. We assume that gender U affects X_2 and Z. In Fig. 4, the left panel shows the causal diagram and the right panel shows the moral graph. In the causal diagram, (Y, X_2) and Z are d-separated by (X_1, U), thus $(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$ holds. Therefore, if the population information of (X_1, U) is available, we can obtain a consistent estimator of the regression coefficients.

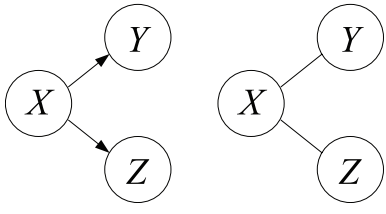

Fig. 3 The left panel shows the causal diagram for Example 1 and the right panel shows its moral graph

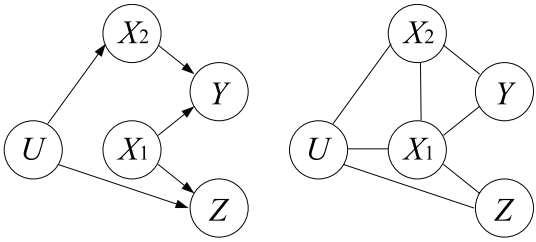

Fig. 4 The left panel shows the causal diagram for Example 2 and the right panel shows its moral graph

Example 3 Let Y be income, X_1 be age, and X_2 be education level. We assume that marital status U is affected by Y, and affects Z. In Fig. 5, the left panel shows the causal diagram and the right panel shows the moral graph. In the causal diagram, (Y, X_2) and Z are d-separated by (X_1, U), thus $(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$ holds. Therefore, if the population information of (X_1, U) is available, we can obtain a consistent estimator of the regression coefficients.

3 Numerical Experiments

In the previous sections, we assumed sampling with replacement and infinite population. In this section, we perform a simulation where sampling is without replacement and population size is finite. Different two sample sizes are used: one value (3500) is typical and the other (10,000) is fairly large in social surveys. The following three examples correspond to the examples of Sect. 2.2.

Example 1 We assume that age X takes values 1, 2, 3, or 4 if the individual is in their twenties, thirties, forties, or fifties, respectively. A population is generated from the distribution

$$\begin{aligned} \Pr(X_1 = 1) = \Pr(X_1 = 2) = \Pr(X_1 = 3) = \Pr(X_1 = 4) = 0.25, \\ Y = [-50(X_1 - 4)^2 + 750 + \varepsilon]_+, \quad \varepsilon \sim N(0, 100^2), \\ \Pr(Z = 1 | X = i) = 0.2i \quad (i = 1, 2, 3, 4), \end{aligned}$$

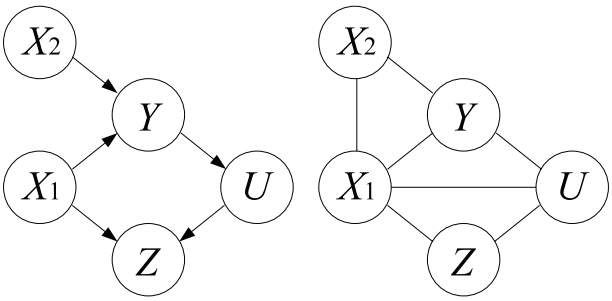

Fig. 5 The left panel shows the causal diagram for Example 3 and the right panel shows its moral graph

where $[x]_+$ is a function returning x if x is positive and 0 otherwise. The size of the population is 100 million. The possibly misspecified regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

is used in the estimation.

To investigate the properties of estimators, sampling with size n was repeated 10,000 times, and averages were calculated from these 10,000 replicates. Table 1 shows the means of the estimates of the regression coefficients. The row labeled PopVal contains the values of the regression coefficients for the population. The row labeled NonAdj contains the mean of the ordinary least square estimates of the regression coefficients based on the data from the respondents. The standard deviation is given in parentheses. The row labeled PostStr contains the mean of the post-stratification estimates of the regression coefficients. The post-stratification estimate is obtained as follows. First, $E(X_2|X_1, U, Z = 1)$, $E(X_2X_2^T|X_1, U, Z = 1)$, $E(Y|X_1, U, Z = 1)$ and $E(X_2Y|X_1, U, Z = 1)$ are estimated by calculating the simple averages for each stratum. Second, each element of (1) is estimated by the technique as in (2). Third, the regression coefficients are estimated by substituting the estimates in (1). The results reveal a tendency for the nonadjusted slope (intersection) estimate to become smaller (larger) than the value for the population. The estimates by the post-stratification are distributed around the value for the population, and the standard deviation becomes smaller as n increases.

Example 2 As in Example 1, we assume that age X_1 takes values 1, 2, 3, or 4. The variable X_2 representing an individual's education level is 1 if the individual is a university graduate and 0 otherwise. Gender U is 1 for a man and 0 for a woman.

A population of size 100 million is generated from the distribution

$$\begin{aligned}
\Pr(U = 1) &= \Pr(U = 0) = 0.5,\\
\Pr(X_1 = 1) &= \Pr(X_1 = 2) = \Pr(X_1 = 3) = \Pr(X_1 = 4) = 0.25,\\
\Pr(X_2 = 1|U = 1) &= 0.4,\ \Pr(X_2 = 1|U = 0) = 0.2,\\
Y &= \begin{cases} [-50(X_1 - 4)^2 + 750 + \varepsilon]_+ & \text{if } X_2 = 1\\ [-30(X_1 - 4)^2 + 500 + \varepsilon]_+ & \text{if } X_2 = 0 \end{cases}, \quad \varepsilon \sim N(0, 100^2),\\
\Pr(Z = 1|X_1 = i, U = j) &= 0.2i - 0.2j + 0.1 \quad (i = 1, 2, 3, 4,\ j = 0, 1).
\end{aligned}$$

The regression model

Table 1 The mean of estimates (Example 1)

n	Estimation method	β_0	β_1
3500	PopVal	199.9	150.0
	NonAdj	260.0 (9.0)	130.0 (2.8)
10000	PostStr	200.3 (8.2)	149.9 (2.6)
	NonAdj	260.0 (5.3)	130.0 (1.6)
	PostStr	200.3 (4.8)	149.9 (1.5)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

is used in the estimation.

To investigate the properties of estimators, sampling with size n was repeated 10,000 times as in Example 1. Table 2 shows the means of estimates of the regression coefficients. PopVal, NonAdj, and PostStr have the same meaning as in Example 1. The row labeled Rak contains the mean of raking estimates of the regression coefficients with auxiliary variable (X_1, U). The raking estimates are obtained by the weighted least square method, where the weight for each respondent is determined by the raking procedure (for example, [10]). The row labeled Rak1 (Rak2) contains the mean of the raking estimates of the regression coefficients with auxiliary variable X_1 (U). In NonAdj, the estimates of β_0 and β_2 have upper biases and the estimate of β_1 has a lower bias. By post-stratification, biases are almost corrected. In Rak, more biases are observed than in post-stratification. In Rak2, bias adjustment has almost no effect.

Example 3 As in Example 2, we assume that age X_1 takes values 1, 2, 3, or 4 and education level X_2 takes values 0 or 1. Marital status U is 1 if the individual is married and 0 otherwise.

A population of size 100 million is generated from the distribution

$$\Pr(X_1 = 1) = \Pr(X_1 = 2) = \Pr(X_1 = 3) = \Pr(X_1 = 4) = 0.25,$$
$$\Pr(X_2 = 1) = 0.4,$$
$$Y = \begin{cases} [-50(X_1 - 4)^2 + 750 + \varepsilon]_+ & \text{if } X_2 = 1 \\ [-30(X_1 - 4)^2 + 500 + \varepsilon]_+ & \text{if} X_2 = 0 \end{cases}, \quad \varepsilon \sim N(0, 100^2),$$
$$\Pr(U = 1 | X_1, Y) = \frac{1}{1 + \exp(-0.004Y + 1)},$$
$$\Pr(Z = 1 | X_1, U) = 0.1i + 0.3j + 0.1 \quad (i = 1, 2, 3, 4,\ j = 0, 1).$$

The regression model

Table 2 The mean of estimates (Example 2)

n	Estimation method	β_0	β_1	β_2
3500	PopVal	125.1	107.9	180.3
	NonAdj	164.8 (8.7)	91.8 (2.8)	212.4 (6.0)
	PostStr	125.0 (10.0)	108.0 (3.2)	180.7 (7.9)
	Rak	126.8 (9.1)	107.3 (3.0)	183.9 (7.4)
	Rak1	130.7 (8.7)	105.8 (2.8)	184.0 (7.3)
	Rak2	164.5 (9.0)	91.8 (2.8)	213.6 (6.1)
10000	NonAdj	164.7 (5.2)	91.8 (1.6)	212.5 (3.5)
	PostStr	124.8 (5.9)	108.0 (1.9)	180.7 (4.6)
	Rak	126.6 (5.3)	107.4 (1.8)	183.9 (4.3)
	Rak1	130.5 (5.1)	105.8 (1.7)	184.0 (4.3)
	Rak2	164.4 (5.3)	91.8 (1.7)	213.7 (3.5)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

is used in the estimation.

To investigate the properties of estimators, sampling with size n was repeated 10,000 times as in Examples 1 and 2. Table 3 shows the means of estimates of the regression coefficients. PopVal, NonAdj, PostStr, Rak, Rak1, and Rak2 have the same meaning as in Example 2. In NonAdj, the estimates of β_0 and β_2 have upper biases and the estimate of β_1 has a lower bias. By post-stratification, biases are almost corrected. In Rak, more biases are left than in post-stratification. In Rak1 the estimates of β_1 and β_2 have small biases, whereas in Rak2 only the estimate of β_0 has a small bias.

4 Application to the SSP-I2010 Survey data

In this section, the "Interview Survey for Stratification and Social Psychology in 2010" (SSP-I2010 Survey) is analyzed. The survey was administered to 3500 Japanese males and females aged 25–59 (at the end of 2009) by the way of face-to-face interviews. The main purpose of this survey was to investigate factors affecting individual stratum identification (economic/social status perception) and public consciousness of economic inequality in Japan. The number of respondents was 1763, yielding a response rate of 50.4%. Detailed information on the survey can be found in SSP Project [12].

In the following analysis, only data for males are used. The number of males in the entire sample is 1717, and the number of observations used in the analysis is 701. The objective variable Y is individual stratum identification (1 (upper) to 10 (lower)). Explanatory variables are age $X_{(1)}$, education level $X_{(2)}$, EGP class categories $X_{(3)}$ which measures occupational prestige of a person [2, 3], and annual income $X_{(4)}$. Education level is divided into three categories (primary, secondary, and higher). Income (ten thousand yen) is categorized as follows: 0, 1–199,

Table 3 Mean of estimates (Example 3)

n	Estimation method	β_0	β_1	β_2
3500	PopVal	− 69.6	113.8	180.0
	NonAdj	− 63.5 (9.4)	105.2 (2.5)	196.5 (5.2)
	PostStr	− 69.6 (10.2)	113.9 (2.6)	180.0 (5.7)
	Rak	− 66.6 (10.0)	112.6 (2.5)	179.9 (5.6)
	Rak1	− 61.0 (9.5)	112.9 (2.4)	179.0 (5.2)
	Rak2	− 70.3 (10.0)	106.2 (2.7)	195.1 (5.6)
10000	NonAdj	− 63.5 (5.6)	105.2 (1.5)	196.5 (3.0)
	PostStr	− 69.6 (6.0)	113.9 (1.5)	180.0 (3.3)
	Rak	− 66.7 (6.0)	112.6 (1.5)	179.9 (3.3)
	Rak1	− 61.0 (5.7)	112.9 (1.4)	179.0 (3.1)
	Rak2	− 70.4 (5.9)	106.2 (1.6)	195.1 (3.3)

Table 4 Variables used in the analysis of the SSP-I2010 Survey

Y	Stratum identification (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
$X_{(1)}$	Age (20s, 30s, 40s, 50s)
$X_{(2)}$	Education level (primary, secondary, higher)
$X_{(3)}$	EGP class (I(higher service), II(lower service), III(routine clerical/sales), IVa(small employer), IVb(independent), V(manual foreman), VI(skilled manual), VIIa(semi-unskilled manual), IVc(farmers/farm managers), without occupation)
$X_{(4)}$	Income (ten thousand yen) (0, 1–199, 200–399, 400–699, 700-999, 1000–1499, 1500–)
$U_{(1)}$	City size (wards, cities with population $\geq 200,000$, cities with population $< 200,000$, towns and villages)
$U_{(2)}$	Occupational status (company president, company executive, self-employed or freelance worker; regular full-time employee; employee dispatched by a temporary employment agency; other irregular employee; family worker; unemployed; not in labor force)
$U_{(3)}$	Marital status (unmarried, married, other)
$U_{(4)}$	House ownership (one's own house, other)
$U_{(5)}$	Household composition (one-person, other)
$U_{(6)}$	Duration of residence (< 5 years, ≥ 5 and < 20 years, ≥ 20 years)

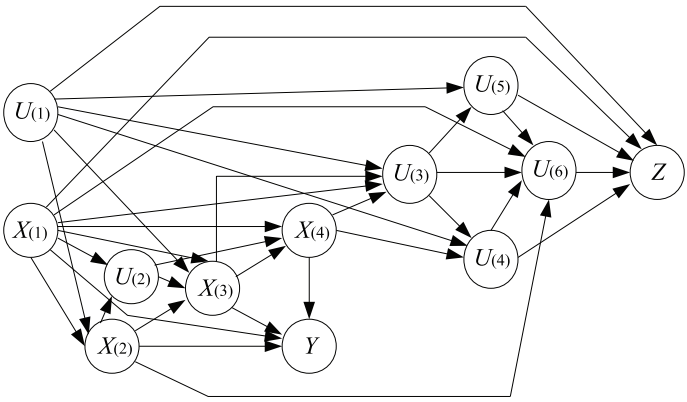

Fig. 6 The causal diagram used in the analysis of the SSP-I2010 Survey data

200–399, 400–699, 700-999, 1000–1499, and more than or equal to 1500. Used auxiliary variables are city size $U_{(1)}$, occupational status $U_{(2)}$, marital status $U_{(3)}$, house ownership $U_{(4)}$, household composition $U_{(5)}$, and duration of residence $U_{(6)}$. Details of the auxiliary variables are shown in Table 4.

We assume causal relationships shown in Fig. 6. From Fig. 6,$(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$ holds, where $X_1 = X_{(1)}, X_2 = (X_{(2)}, X_{(3)}, X_{(4)})$ and $U = (U_{(1)}, \ldots, U_{(6)})$. Population information for $X_{(1)}$ and each $U_{(i)}$ can be obtained from the 2010 Population Census of Japan (Ministry of Internal Affairs and Communications, [9]).

Estimated regression coefficients are shown in Table 5, where Age (20s), Education level (primary), EGP class (without occupation), and Income (0) are reference categories. NonAdj is the ordinary least square estimate of the regression. Rak is the weighted least square estimate where the weights are determined by the raking procedure with auxiliary variable U. In Table 5, we can see a tendency that the absolute values of estimated regression coefficients for education level become smaller in Rak while the absolute values of estimated regression coefficients for EGP class become larger.

5 Summary

In this study, we considered a nonresponse bias adjustment problem in regression analysis. If a linear model assumed in regression analysis is not correct, the least square estimator may be biased. We established a sufficient condition that allows one to obtain a consistent estimator. Whether the condition holds is determined by the causal diagram.

For the analysis, we first create a causal diagram based on prior knowledge. Next, we find auxiliary information U satisfying the conditional independence$(Y, X_2) \perp\!\!\!\perp Z \mid (X_1, U)$, where population information of (X_1, U) is available.

Table 5 Estimated regression coefficients

	NonAdj	Rak
(Intercept)	7.80	7.85
Age (30s)	0.14	0.02
Age (40s)	0.41	0.38
Age (50s)	0.20	0.07
Education level (secondary)	− 0.63	− 0.44
Education level (higher)	− 0.94	− 0.76
EGP class (I)	− 0.87	− 1.02
EGP class (II)	− 0.86	− 0.99
EGP class (III)	− 0.64	− 0.74
EGP class (IVa)	− 1.02	− 1.25
EGP class (IVb)	− 0.84	− 0.98
EGP class (V)	− 0.50	− 0.52
EGP class (VI)	− 0.22	− 0.28
EGP class (VIIa)	− 0.10	− 0.19
EGP class (IVc)	− 0.12	− 0.50
Income (1–199)	− 0.48	− 0.26
Income (200–399)	− 0.71	− 0.63
Income (400–699)	− 0.96	− 0.98
Income (700–999)	− 1.84	− 1.84
Income (1000–1499)	− 1.75	− 1.89
Income (1500–)	− 2.62	− 2.15
Coefficient of determination	0.31	0.35

Finally, the weighted least square estimate is calculated, where the weight for each respondent is obtained by the calibration procedure [11] with auxiliary variables (X_1, U).

In this study, linear regression models were analyzed. In real data analysis, a generalized linear regression model such as a logistic regression model is often used. The sufficient condition established in this study is also valid for the generalized linear regression model. Thus, the condition can be widely applied to analyze survey data.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 15K00043.

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Bethlehem J, Cobben F, Schouten B (2011) Handbook of nonresponse in household surveys. Wiley, New Jersey
2. Erikson R, Goldthorpe JH, Portocarero L (1979) Intergenerational class mobility in three Western European societies. Br J Sociol 30:415–441
3. Ganzeboom HBG, Treiman D (1996) Internationally comparative measures of occupational status for 1988 international standard classification of occupations. Soc Sci Res 25:201–239
4. Groves RM, Couper MP (1998) Nonresponse in household interview surveys. Wiley, New York
5. Groves RM, Dillman D, Eltinge JL, Little RJA (eds) (2002) Survey nonresponse. Wiley, New York
6. Ibrahim JG, Chen M-H, Lipsitz SR, Herring AH (2005) Missing data methods for generalized linear models: a comparative review. J Am Stat Assoc 100:332–346
7. Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford
8. Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, Hoboken
9. Ministry of Internal Affairs and Communications (2014) 2010 Population Census of Japan. http://www.e-stat.go.jp/. Accessed 19 Feb 2020
10. Oh HL, Scheuren FJ (1983) Weighting adjustment for unit nonresponse. In: Madow WG, Olkin I, Rubin DB (eds) Incomplete data in sample surveys, vol 2. Academic Press, New York
11. Särndal CE, Ludström S (2005) Estimation in surveys with nonresponse. Wiley, Chichester
12. SSP Project (2013) Codebooks and basic summary tables of the SSP-I2010 Survey, Osaka: Author (in Japanese)
13. van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge