MODELING THE FACTORS THAT INFLUENCED THE OUTCOMES OF INTERVIEW OF THE JULY 2006 LABOR FORCE SURVEY IN THE NATIONAL CAPITAL REGION: A LOGISTIC REGRESSION APPROACH

by

Maribel L. Bernardo

For additional information, please contact:

Author's name: Maribel L. Bernardo
Designation: Statistician
Affiliation: National Statistics Office - MIMAROPA
Address: J.P. Rizal Street, Barangay Camilmil, Calapan City, Oriental Mindoro
Tel. no.: +6343-4410398
E-mail: m_bernardo@yahoo.com
MODELING THE FACTORS THAT INFLUENCED THE OUTCOMES OF INTERVIEW OF THE JULY 2006 LABOR FORCE SURVEY IN THE NATIONAL CAPITAL REGION: A LOGISTIC REGRESSION APPROACH

by

Maribel L. Bernardo and Florante C. Varona

ABSTRACT

The main objective of the study is to identify significant factors that influenced the outcomes of the interview of Labor Force Survey (LFS) conducted by the National Statistics Office (NSO). This study used the characteristics of supervisors and enumerators based from the administrative records of NSO, as well as the characteristics of the household head of sample households of LFS to model the factors that influenced the outcomes of interview.

Since July 2006 and January 2007 rounds of LFS used the same sets of sample households, the characteristics of responding households were taken from the successful interviews of the July 2006 LFS, while the characteristics of non-responding households were taken from the unsuccessful interviews of the July 2006 LFS, but successfully interviewed during the January 2007 LFS.

Logistic regression was used to model the factors that influenced the outcomes of interview. The significant differences on characteristics of responding and non-responding households were examined using t-tests and chi-square tests.

Findings of the study revealed that the significant characteristics that contribute to the model are coordination of supervisor by providing letter to the barangay chairperson/official prior to the conduct of the survey, single marital status of the enumerator, age of the household head, and higher education for the household head.

It is recommended that factors with significant positive influenced on the outcomes of the interview be given consideration during field operation of survey or census.

I. Introduction

In every census and survey conducted, errors in the estimates are present either due to sampling or non-sampling. Error due to non-sampling for example can be attributed in the case when the enumerator has no control on how the respondent understands the questions during the interview. Factors such as the geographic location of sample household, ways on how the enumerator approached the respondent, time, length and topic of the interview and weather conditions at the time of visit can also affect the success of the surveys. The target respondents who did not cooperate during the interview resulted to nonresponse cases. For the census results, this would mean lower count which undermines the essence of a complete enumeration. Similarly, for surveys, this also means lesser or reduced number of samples. If this matter is taken for granted, the bias contributed by nonresponse error in the estimates might be larger compared to the errors due to sampling. It is, therefore, important to minimize nonresponse cases in any surveys undertaken because this reduces the number of elements on which to base the estimates for analyses.

This study was conducted in order to determine the significant factors that contributed to the unit nonresponse (household that were not successfully interviewed) and unit response (households that were successfully interviewed) using the results of the July 2006 and January 2007 rounds of the Labor Force Survey (LFS) in the National Capital Region (NCR) by the NSO. The study utilized selected personal information of the
supervisors and enumerators involved in these two survey rounds. Respondent households who were successfully interviewed in July 2006 LFS round were used in the study. The characteristics of the sample households who were not successfully interviewed during the July 2006 LFS round were taken from the January 2007 LFS round from which they were successfully interviewed. In this study, the characteristics of the head of the sample households such as age, sex, highest grade completed, employment status, major type of occupation, broad sector of industry and household size were considered. Moreover, for supervisors and enumerators, their age, sex, marital status, highest grade completed, experience as census interviewer/supervisor in terms of number months served and their usual residence were utilized. Also included as supervisor’s characteristics are its issuance of map to interviewers, provision of advance letter to the sample households and barangay chairperson/official prior to the conduct of the survey.

The effects of the nonresponse in the overall quality of data can be initially exemplified in the following discussions:

Mr. Groves and Couper (1998) showed in Figure 1 the four different cases of statistical impact of nonresponse on survey estimates. For purposes of discussions of the four different cases below in Figure 1 referred to as the respondents mean and, \( \bar{y}_m \) mean of the nonrespondents.

Case (a) in Figure 1 reflects a high response rate survey and one in which the nonrespondents have a distribution of \( y \) values quite similar to that of the respondents. This case will produce the lowest bias.

Case (b) shows a very high nonresponse rate. That is, the area under the respondent distribution is about 50 percent greater than the nonresponse rate of the nonrespondent. Similar with case (a), the values of \( y \) in the nonrespondents are similar to those of the nonrespondents. Thus, the respondent mean has low bias due to nonresponse.

Case (c) is similar with case (a) low nonresponse survey. However, in this case, the nonrespondents tend to have much higher values than the respondents. This means that the respondent mean underestimates the full population mean. However, the size of the bias is small because of the low nonresponse rate of about 5% or so.

Case (d) is the most perverse. It exhibits a large group of nonrespondents, who have much generally higher values on \( y \) compared to the values of the respondents. This is a case of large nonresponse bias.
II. Purpose and Objectives

The findings of the study can be used as basis in developing a better strategy and approach to census and survey taking designed to elicit cooperation from households to respond to interviews. In effect, the non-sampling error due to nonresponse in LFS will be lessened, thus, the quality of statistics derived from this survey can be further improved.

III. Methods

This study used the descriptive research method in analyzing the factors that affect the interview status of sample households of the July 2006 LFS. The factors considered were characteristics of supervisor, enumerator and household head.

The study used secondary data from the results of the LFS with reference period of July 2006 and January 2007. The July 2006 and January 2007 were selected as reference periods because the sample households in these rounds were completely overlapping. This
means that the two rounds used the same sets of sample households. This is because the 2007 Family Income and Expenditures Survey (FIES) of NSO, which was conducted in two visits using the same sample households, was a rider survey of the 2006 LFS and 2007 LFS. The 2007 FIES employed a shuttle type questionnaire wherein the same respondent was interviewed in two different periods. The first visit of 2007 FIES was in July 2006 while the second visit was in January 2007.

The characteristics of the household head of responding households such as age, sex, marital status, highest grade completed, employment status, major type of occupation, broad sector of industry, and household size were based on the 2006 LFS while the characteristics of the household head of nonresponding households were based on the January 2007 LFS. For nonresponding households in the 2006 LFS and 2007 LFS, these were excluded in the analysis due to lack of needed information on the characteristics of these households. The characteristics of enumerators and supervisors were taken from the administrative records of NSO. These characteristics are as follows:

For enumerators: age, sex, marital status, highest grade completed, experience as census interviewer in terms of months and residence of enumerator vis-à-vis area of assignment.

For supervisors: age, sex, marital status, highest grade completed, experience as census interviewer in terms of months, residence of supervisor vis-à-vis area of assignment, whether or not provided enumeration map to the enumerator, whether or not provided advance letter to the sample household and whether or not provided advance letter to the barangay chairman/official.

The variables used in this study were taken from the complete raw data of July 2006 and January 2007 LFS. Extraction of the variables and some demographic characteristics of the household head that are specific for the study was done with the use of the Census and Survey Processing (CSPro) software. This is the main software that NSO is using for machine processing of the results of its censuses and surveys.

The output file of CSPro can be read by other softwares such as SPSS, Stata, and SAS when converted into the data specification of the said softwares.

Statistical Treatment of Data
Descriptive Analysis, Chi-Square test, and T-Test

Descriptive statistics such as frequencies, percentages, means and medians were utilized to describe the samples of this study.

To answer the first research question on whether or not there are significant differences between responding and nonresponding households in terms of the selected household head characteristics, t-test and chi-square test were applied.

T-test was used in comparing the mean age and household size of the responding households against the nonresponding households since these variables are in ratio scale. If each group is assumed to be normally distributed and that the population variances are equal \((\sigma_1^2 = \sigma_2^2)\), a t test with \(n_1 + n_2 - 2\) degrees of freedom can be used to determine if there is significant difference between the means of the two groups. The t statistic is given as
\[ t_{n_1+n_2-2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \]  

with \( n_1 + n_2 - 2 \) degrees of freedom.

where,

\[ S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2} \]

\( S_p^2 \) = pooled variance of the two groups

\( \bar{X}_1 \) = sample mean of group 1

\( S_1^2 \) = sample variance of group 1

\( n_1 \) = sample size of group 1

\( \bar{X}_2 \) = sample mean of group 2

\( S_2^2 \) = sample variance of group 2

\( n_2 \) = sample size of group 2

On the other hand, the t-statistic assuming normality of the two groups and unequal variance is

\[ t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \]

with degrees of freedom

\[ v = \frac{(S_1^2 / n_1 + S_2^2 / n_2)^2}{\frac{(S_1^2 / n_1)^2}{n_1-1} + \frac{(S_2^2 / n_2)^2}{n_2-1}} \]

In this study, group 1 is referred to as the responding households and group 2 as the nonresponding households.

The hypothesis that the two groups do not differ is rejected when the t statistic falls on the rejection region that is, \( t < -t_{\alpha/2} \) or \( t > t_{\alpha/2} \) where \( \alpha = 0.05 \) level of significance.

The nonparametric Mann-Whitney U test was also considered to validate the results of the T-test. Mann-Whitney U test does not require normal distribution of data.

Levene’s test is used to test if k groups of samples have equal variances. K groups of samples with equal variances are homogeneous. Some statistical tests, for example analysis of variance, assume that variances are equal across groups of samples. The Levene test can be used to verify that assumption of homogeneity.

Chi-square test was applied to know if the household heads of the responding and nonresponding households are the same in terms of sex, marital status, highest grade completed, employment status of household head, major type of occupation, and broad sector of industry.
The Chi-square ($X^2$) is given as follows:

$$X^2 = \sum_{i=1}^{rc} \frac{(o_i - e_i)^2}{e_i}$$

where,
- \(r\) refers to the number of rows.
- \(c\) refers to the number of columns.
- \(o_i\) refers to the observed frequency of the \(i\)th cell.
- \(e_i\) refers to the expected frequency of the \(i\)th cell.

Expected frequencies are computed as

$$e_i = \frac{(column\_total) \times (row\_total)}{grand\_total}$$

where the summation extends overall \(rc\) cells in the \(r \times c\) contingency table. If $X^2 > X^2_a$ with \(v=(r-1)(c-1)\) degrees of freedom, reject the null hypothesis that the characteristics of the household head of responding and nonresponding households have no difference at \(\alpha\) level of significance. Otherwise, accept the null hypothesis.

**LOGISTIC REGRESSION**

The second research question used logistic regression in identifying the factors that influenced the interview status of the July 2006 LFS.

Logistic regression is a specialized form of regression formulated to predict and explain a binary (two-group) categorical variable or an ordinal variable rather than a metric dependent measure. It is also known as logit analysis. The model yields regression-like coefficients that indicate the relative impact of each predictor variable.

The logistic regression model gives the probability that a sample household will respond or will not respond in the July 2006 LFS.

Logistic regression is similar to multiple regression in the sense that both techniques can be used to describe the relationship of several explanatory variables and a response variable. The analysis and interpretation of logistic regression are also quite similar to that of least squares regression. Thus, the techniques for model selection and hypothesis testing in logistic regression follow the pattern associated with least squares regression. However, in logistic regression, the dependent variable can be a dichotomous variable or an ordinal variable.

Ordinary least squares (OLS) regression is not appropriate when the dependent variable is not interval scaled because of the following reasons:

1. predicted probabilities might be less than 0 or greater than 1;
2. variance of the error term is not constant, hence some explanatory variables that are significant may be excluded from the model;
3. since the dependent variable is binary, the appropriate distribution for the error term is the binary distribution hence residuals will not be normally distributed; and
4. relationship between the conditional probability that $Y=1$ and a continuous $X$ is usually not linear but is S-shaped.
Assumptions:

Logistic regression requires categorical or non-metric dichotomous dependent variable, and this makes it equivalent to two-group discriminant analysis. However, logistic regression is more preferred than discriminant analysis in two-group case as it does not require the strict assumptions of multivariate normality and equal variance.

The logistic regression model

The probability that a sample household will respond in the LFS, \( P(Y = 1) \), can be estimated using

\[
P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)},
\]

where \( X_1, \ldots, X_n \) are the independent variables and \( \beta_0, \ldots, \beta_n \) are the regression coefficients. In this study, the independent variables are the characteristics of the household heads, enumerators and field supervisors. Specifically, the variables and their categories names and the variable names in the dataset are as follows:

1. supervisor’s characteristics
   - age
   - sex
   - marital status
   - highest grade completed
   - number of months of experience as census interviewer
   - residence of supervisor vis-à-vis area of assignment
   - provided enumeration map to the enumerator
   - provided of advance letter to sampled household
   - provided of advance letter to the barangay chairman

2. enumerator’s characteristic
   - age
   - sex
   - marital status
   - highest grade completed
   - number of months of experience as census interviewer
   - residence of enumerator vis-à-vis area of assignment

3. household head’s characteristics
   - age
   - sex
   - marital status
   - highest grade completed
   - employment status
   - type of occupation
   - broad sector of industry
   - household size

Automatic recoding of categorical variables is being done by SPSS. For the dependent variable or outcome of interview, households who responded were coded as 1 and 0 for households who did not respond. The reference category for categorical variables was based from the highest frequency.
The \( \beta \)'s in the logistic regression are estimated through the process of maximum likelihood estimation. The logit iterative estimation procedure maximizes the logarithm of the likelihood function. At iteration 0, the log likelihood describes the fit of a model including only the constant. The last log likelihood describes the fit of the final model.

Instead of modeling the probability that a sample household will respond in the LFS, that is, \( P(Y = 1) \), the odds ratio can be modeled, where odds ratio is the ratio of the probability that some event will occur to the probability that the same event will not occur. The odds ratio, which takes values between 0 and 1 is given by

\[
\text{Odds Ratio} = \frac{P(Y = 1)}{P(Y = 0)} \quad \text{or} \quad \text{Odds Ratio} = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)
\]

The odds ratio can be expressed in terms of its logarithm which is called logit model. Hence, the logit model or log odds ratio can be expressed as \( Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \).

**Goodness-of-fit test for the estimated model**

1. \( \chi^2 \)-test

An overall \( \chi^2 \)-test evaluates the null hypothesis that all coefficients in the model, except the constant, equal to zero. It is a test of the observed against expected number of responses using cells defined by the covariate patterns. Two observations are said to share the same covariate pattern if the independent variables for the two observations are identical.

   If the p-value of the \( \chi^2 \)-test is less than 0.05, then the explanatory variable have a significant effect on the odds of being able to respond in the LFS.

2. Pseudo \( R^2 \)

Logit displays a pseudo \( R^2 \). Pseudo \( R^2 \) provides a quick way to describe or compare the fit of different models for the same dependent variable. The "pseudo \( R^2 \)" for logit model was computed as follows:

\[
pseudo R^2 = \frac{-2LL_{\text{null}} - (-2LL_{\text{model}})}{-2LL_{\text{null}}},
\]

The higher the "pseudo \( R^2 \)" the better.

3. Method of classification table

There are other ways to assess the overall model fit aside from the \( \chi^2 \)-test and the "pseudo \( R^2 \)". In addition to the test mentioned earlier, this study also made use of the method of classification matrices to assess the predictive accuracy in terms of group membership.
Sensitivity is the fraction of observed positive-outcome cases that are correctly classified. Specificity is the fraction of observed negative-outcome cases that are correctly classified. The overall "correctly classified" rate was also computed.

Testing for significance of the regression coefficients

In multiple regression, the t-value is used to assess the significance of each regression coefficient. Logistic regression uses a different statistic, the Wald statistic. It provides the statistical significance of each estimated coefficient so that hypothesis testing can occur just like the t-test in multiple regression.

Interpreting the coefficients

The estimated coefficients, $\beta_0, \beta_1, \ldots, \beta_n$, are actually measures of the changes in the ratio of the probabilities called odds ratio. They can be expressed in logarithms, as defined earlier, and are transformed back by taking the antilog of the value so that their relative effect on the probabilities is assessed more easily. A positive coefficient increases the probability of occurrence of the event, whereas a negative value decreases the predicted probability of no occurrence of the event.

IV. Findings/Results

1. The mean age of household heads of responding households of the January 2007 of Labor Force Survey (LFS) in National Capital Region (NCR) was 45.81, lower than the 50.56 mean age of household heads of nonresponding households. The mean age of heads of responding households was significantly different from the mean age of heads of nonresponding households.

2. The average household size of responding households of the January 2007 LFS in NCR was 4.86, which was lower than the average households size of 5.06 of the nonresponding households. There was no significant difference of the average household size between the responding and nonresponding households.

3. There was no significant difference in the proportion of responding and nonresponding households of the January 2007 LFS in NCR in terms of sex, marital status, employment status, and major sector of industry of household heads.

4. There was a significant difference in the proportion of responding and nonresponding households of the January 2007 LFS in NCR in terms of highest grade completed and major occupation of household head.

5. Among supervisor’s characteristics, providing letter to barangay captain/official was a significant predictor of the interview status of the household. The predicted probability that the household will respond using this explanatory variable is given by the logistic regression equation,

$$P(Y = 1) = \frac{\exp(3.022 + 0.638S_{LBC})}{1 + \exp(3.022 + 0.638S_{LBC})}$$

6. It was surprising to note that in NCR, households in barangays where supervisors did not provide letter to the barangay captain/official were more likely to respond than households in barangays where supervisors provided letter to the barangay.
captain/official. The odds of responding when barangay captain is not provided with letter was almost twice (1.89) the odds of responding when providing letter to the barangay captain/official.

7. Among enumerator’s characteristics, marital status was a significant predictor of the interview status of the household. The predicted probability that the household will respond with marital status as predictor is given by the logistic regression equation,

\[
P(Y = 1) = \frac{\exp(2.994 + 0.315E_{MS_{Re}})}{1 + \exp(2.994 + 0.315E_{MS_{Re}})}
\]

8. Single enumerators were more likely to get response from the household than the rest of the enumerators where the odds of single was 1.37 times that of the rest of the marital status.

9. Among the household head’s characteristics, age and highest grade completed of the household head were significant predictors of the interview status of the household. The predicted probability that the household will respond using age and highest grade completed of the household head as explanatory variables is given by the logistic regression equation,

\[
P(Y = 1) = \frac{\exp(4.704 - 0.026_{age} + 0.485_{HGC_{Rec_{Head}}(1)} - 0.732_{HGC_{Rec_{Head}}(2)} - 0.889_{HGC_{Rec_{Head}}(3)})}{1 + \exp(4.704 - 0.026_{age} + 0.485_{HGC_{Rec_{Head}}(1)} - 0.732_{HGC_{Rec_{Head}}(2)} - 0.889_{HGC_{Rec_{Head}}(3)})}
\]

10. Older household heads were less likely to respond than younger household heads. The odds of responding for older household heads was 97.4 percent of the odds of responding for younger household heads.

11. As compared to household heads who had reached high school level, household heads with no grade completed or had reached elementary level were more likely to respond with an odds of 1.624. However, household heads who had reached college were less likely to respond than heads who reached high school with an odds of 0.481 for college undergraduate and 0.411 for college graduate.

12. Combining all the characteristics of the supervisor, enumerator and household head that were found to be significant predictors of interview status, the predicted probability that the household will respond is given by

\[
P(Y = 1) = \frac{\exp(4.465 - 0.517S_{LSC} + 0.336E_{Mls_{Rec}} - 0.025_{age} + 0.589_{HGC_{Rec_{Head}}(1)} - 0.759_{HGC_{Rec_{Head}}(2)} - 0.879_{HGC_{Rec_{Head}}(3)})}{1 + \exp(4.465 - 0.517S_{LSC} + 0.336E_{Mls_{Rec}} - 0.025_{age} + 0.589_{HGC_{Rec_{Head}}(1)} - 0.759_{HGC_{Rec_{Head}}(2)} - 0.879_{HGC_{Rec_{Head}}(3)})}
\]

V. Conclusion, Recommendation, and Implications

Based on the findings of the study, a model using logistic regression can be used to analyze the outcome of interview using the data of the July 2006/January 2007 LFS of NSO. The following variables were found to have significant influence on the outcome of the interview of the January 2007 LFS in NCR:

- Supervisor providing letter to the barangay chairman/official;
- Marital status of the enumerator;
• Age of the household head; and
• Highest educational attainment of the household head

It is recommended that factors with significant influence on the outcome of the interview be given consideration during recruitment and field operations whether in survey or censuses.

1. In hiring for enumerators for instance, preference should be given to those who have single marital status.

2. For households whose heads are older and with higher educational attainment, response rate can be improved through a better and appropriate field operation strategies such as, tailor-made interview techniques, and extensive information campaign drive. Since heads who are older and heads who have attained college level were found to have lower tendency to respond, date and time of interview should be well planned and use of Self-Administered Questionnaire (SAQ) is recommended. Inasmuch as these heads have pursued higher education, it is most likely that they are employed, hence they are not at home most of the time.

3. Since results of this study pertains only to NCR, results for other regions can be obtained using the same techniques used in this undertaking.

4. The results of the study can be verified using census data, which can provide more cases of nonrespondents and sufficient representation at the regional level.