# Nonresponse Bias in the Survey of Occupational Injuries and Illnesses October 2013

Erin M. Huband, Patrick Bobbitt
Bureau of Labor Statistics, 2 Massachusetts Ave NE Room 3160, Washington, DC 20212

## Abstract

The Bureau of Labor Statistics Survey of Occupational Injuries and Illnesses collects data to estimate the counts and rates of work-related injuries and illnesses. Participation by private sector employers is mandated by the Occupational Safety and Health Act of 1970. For state and local government establishments, however, state laws determine whether the survey is mandatory. While private sector response rates reflect the mandatory nature of the survey, response rates for states in which public sector response is voluntary are low. To determine whether the survey suffers from bias attributable to non-response, government units were classified as either “likely” or “unlikely” respondents using a logistic regression model. Counts and rates of injuries and illnesses for these groups were then compared to provide an indication of potential nonresponse bias. This paper describes the methodology used for this analysis and presents some preliminary results from the analysis.

**Key Words:** nonresponse bias, logistic regression

## 1. Overview

### 1.1 Survey Scope

The Survey of Occupational Injuries and Illnesses (SOII), administered by the Bureau of Labor Statistics (BLS), provides annual information on the rates and counts of work-related injuries and illnesses, and how these statistics vary by incident, industry, geography, occupation, and other characteristics. Each yearly sample of workplaces selected by the BLS consists of approximately 240,000 establishments. SOII data are solicited from employers having 11 or more employees in agricultural production, and from employers of all sizes in all other industries. Starting with survey year 2008, SOII also collected data from state and local government establishments to provide estimates of occupational injuries and illnesses among government workers for the nation and each state. In 2011, the portion of establishments that were government was only 1.3%; of the national employment, only 6.2% was in governments. Prior to 2008, state and local government injury and illness estimates were available for only a selection of states and at varying levels of detail. Self-employed persons are not considered to be employees under the 1970 act. Private households, the United States Postal Service, and federal government workers are also out of scope for the SOII. Most SOII data are directly collected from employers, except for data in mining and railroads in which case the data come from the Mine Safety and Health Administration of the U.S. Department of Labor and the Federal Railroad Administration of the U.S. Department of Transportation.

### 1.2 Survey Sample Design

Because the SOII is a federal-state cooperative program and the data are designed to meet the needs of the states, an independent sample is selected for each state. The survey uses a stratified sample design, with strata of state, ownership (whether private, state government, or local government), industry, and size class (a grouping defined by the establishment's average annual employment). All establishments in the largest size class are selected with certainty. Sample sizes are determined by the participating states based on budget constraints. The optimal allocation procedure used by the SOII distributes the sample to the industries in a manner intended to minimize the variance of the total number of recordable cases in the universe or, alternatively, the incidence rate of recordable cases in the universe. In strata with higher variability of the data, a larger sample is selected.

### 1.3 Survey Estimation

Data collected for the SOII are used to tabulate estimates for two separate data series: annual summary (industry-level) estimates and more-detailed case circumstance and worker characteristic estimates for cases that involved days away from work. Part of the estimation process involves weighting sample units and cases to represent all injuries and illnesses from units on the frame from which the sample was selected.

### 1.4 Survey Nonresponse

A SOII response rate analysis from 2011 showed that for the years 2003 – 2010, overall response rates slowly declined from 94% to 90%. It was found that private industry response rates didn't vary much from 91%, but response rates for state and local governments dropped dramatically starting in 2008. This is the year when the BLS began collecting government data for all states. Previous to this, it had only been collecting state and local government data from a small number of states.

Even though all states and their government establishments are now being surveyed, there remain some states where reporting the government injury and illness data is voluntary. When we looked only at those states where reporting data for the state and local governments is not required by law, response rates for governments are low (between 30% — 50% in 2010). Though overall response rates are not low enough to trigger an Office of Management and Budget-mandated nonresponse bias analysis, those for public sector data in voluntary states are.  It is in this case that nonresponse bias was studied.

### 1.5 Nonresponse Bias Analysis

As part of this nonresponse bias analysis, we assessed what factors influenced a respondent actually responding. Using a logistic regression model, we used the responses of those least likely to respond as a proxy for those that did not respond. Comparing likely responders to non-likely responders allowed us to measure any nonresponse bias.

## 2. Methodology

### 2.1 Overview

Data for this analysis were compiled from the database of all sample units from the 2009 – 2011 SOII. There are 23 states for which public sector reporting is voluntary. The entire samples (including government and private sector units) for these 23 states comprise 280,016 units (37%) of all 755,545 SOII sample units for the three-year period. Nonviable units (units that had gone out of business or were out of scope for the survey, that were reported for in more than one way, or for which good addresses were not available) and units from the private sector were removed from the dataset, leaving 19,067 observations. Descriptive information such as how many employees each unit had, each unit's state, each unit's industry, number of injury and illness case for each unit, and the sampling weight for each unit was also included in the dataset. Each unit was identified as either a respondent or nonrespondent, based on whether the unit responded to the SOII. Because a unit's status as a respondent is a binary variable, logistic regression was a good choice.

Like linear regression, logistic regression uses input (or independent) variables. Unlike linear regression, logistic regression uses the independent variables to predict the probability of the outcome (or dependent variable) occurring. Once we predicted whether a unit will respond, we used the predictive probabilities from the model to predict if similar units in the future will respond to the SOII. Having information about whether units are likely to respond helped us build more efficient allocation algorithms. After using the model to predict the units to be respondents or nonrespondents, we compared the two groups to see if any bias exists.

## 2.2 Logistic Regression Model

It was shown that geographical region in which the unit exists, size class of the unit (five groups based on the number of employees each unit has), and industry sector have an effect on whether a unit will respond (Huband, 2010). The set of independent variables used for this analysis was as follows:

- Ownership (whether in state or local government)
- Region (six geographical regions plus one for the states whose data are collected by staff in the national office)
- Size class (five levels depending on the number of employees)
- Supersector (eleven different categories of industries)
- Weight class (five levels depending on the sample weight of the establishment)
- Certainty (whether the unit was selected to be in the sample with probability 1)
- Interaction between region and supersector

From the SAS output obtained after running the logistic regression model, it gave a likelihood ratio chi-square test statistic value of 1960.9431. This verified that our model fits significantly better than a model with no independent variables. The other test statistics, Score and Wald (1822.8242 and 1468.8408 respectively), also indicated significance of the model. Among all the initial independent variables, weight class was the only one found to not add to the model's fit; therefore, it was dropped in the stepwise logistic regression. The point estimates and standard errors for the independent variables can be found in Appendix A.

The following table shows the independent variables significance test results, each with a significant (at $\alpha = 0.05$) outcome.

**Table 1:** Independent Variables and
Corresponding Test and Significance Statistics

| Independent Variable | Degrees of Freedom | Wald Chi-Square | Significance Level |
|---|---:|---:|---:|
| Ownership | 1 | 150.1707 | < 0.0001 |
| Region | 5 | 11.1840 | 0.0479 |
| Size class | 4 | 413.9952 | < 0.0001 |
| Supersector | 10 | 38.3634 | < 0.0001 |
| Certainty | 1 | 35.2425 | < 0.0001 |
| Region * Supersector | 47 | 180.5169 | < 0.0001 |

## 2.3 Classification Test

After the model was built and found to fit the data well, we had to see how well the model classified the units as respondents. To do this, a classification test was run. The first step in running a classification test is to split the data into two groups: one used for creating the model, the other used for testing the model. The dividing of the data was done by simple random sampling, with 90% of the dataset being used for the model building, and the remaining 10% used to test the model. We used this 90/10 split to ensure that we had enough data on which to build the model, while the test data had enough observations to estimate the model's performance.

The model built on the 90% was then applied to the randomly-selected 10%. Based on the new model, units were classified as "likely" respondents if the probability of responding was greater than some cut-off value (which ranged from 0.20 to 0.90). If the probability of responding was less than the cut-off value, the establishment was classified as an "unlikely" respondent.

Next, the results of the classification test were summarized in confusion matrices for each of the eight cut-off values.

| | Actual Respondent | Actual Nonrespondent |
|---|---|---|
| Predicted Respondent | True Positives | False Positives |
| Predicted Nonrespondent | False Negatives | True Negatives |

Using the true positives, false positives, and false negatives, the precision (the proportion of those predicted as respondents that actually responded), recall (the proportion of actual respondents correctly predicted to be respondents), and $F_1$ scores were calculated. The $F_1$ score (or F-measure) combines precision and recall.

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

These three statistics helped determine our cut-off value. In this case, the $F_1$ score reaches a maximum at the cut-off of 0.40. The table below shows these summary statistics (as well as values from the confusion matrices) for each of the eight cut-off values.

**Table 2:** Possible Cut-off Values with Confusion Matrix Statistics

| Cut-off Value | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 |
|---|---|---|---|---|---|---|---|---|
| True Positives | 766 | 702 | 561 | 343 | 133 | 59 | 43 | 37 |
| False Positives | 1033 | 855 | 496 | 271 | 81 | 16 | 5 | 1 |
| False Negatives | 15 | 79 | 220 | 438 | 648 | 722 | 738 | 744 |
| True Negatives | 93 | 271 | 630 | 855 | 1045 | 1110 | 1121 | 1125 |
| Precision | 0.426 | 0.451 | 0.531 | 0.559 | 0.621 | 0.787 | 0.896 | 0.974 |
| Recall | 0.981 | 0.899 | 0.718 | 0.439 | 0.170 | 0.076 | 0.055 | 0.047 |
| $F_1$ | 0.594 | 0.601 | 0.610 | 0.492 | 0.267 | 0.138 | 0.104 | 0.090 |

The $F_1$ score is at a maximum when the cut-off value is 0.40, but precision and recall reach their maximums at opposing values of the cut-offs. To find the value that made sense for all three statistics, we treated the precision and recall as means of Bernoulli random variables and created confidence intervals around them. For cut-off values less than 0.40, the precision values are not significantly different from each other. For cut-off values greater than 0.60, the recall values are not significantly different from each other. But for cut-off values 0.40 and 0.50 precision does not differ significantly, while the recall does. It is for this reason that 0.40 was selected as the cut-off value. While it might seem like 0.50 would be the best choice, we found that when using the cut-off of 0.40, the predicted response rate mimicked what we found in the data, thereby solidifying our choice.

## 2.4 Applying the Classification

Once the cut-off value was chosen, the original model was applied to the original dataset, and the probability of response was calculated for each respondent. If that probability was greater than 0.40, then the unit was labeled a “likely” responder. If, however, the probability was less than 0.40, then the unit was labeled an “unlikely” responder.

## 2.5 Comparison

Once we had the predictions for which of our establishments would respond or not, we compared the means of original sampling weights, raw counts of total injuries and illnesses, weighted counts of total injuries and illnesses, and injury and illness rates for the two groups (likely and unlikely respondents) and found that for each year, they differed significantly at the $\alpha = 0.05$ level, indicating that there is potential for nonresponse bias. The comparisons were made using *t*-tests across many different levels.

From the charts below, we can see that mean case counts and mean case rates differ for likely respondents and unlikely respondents. And comparing the two charts, we see that the direction of the differences changes between counts and rates; case rates take into account the hours worked, which could explain some of the difference.

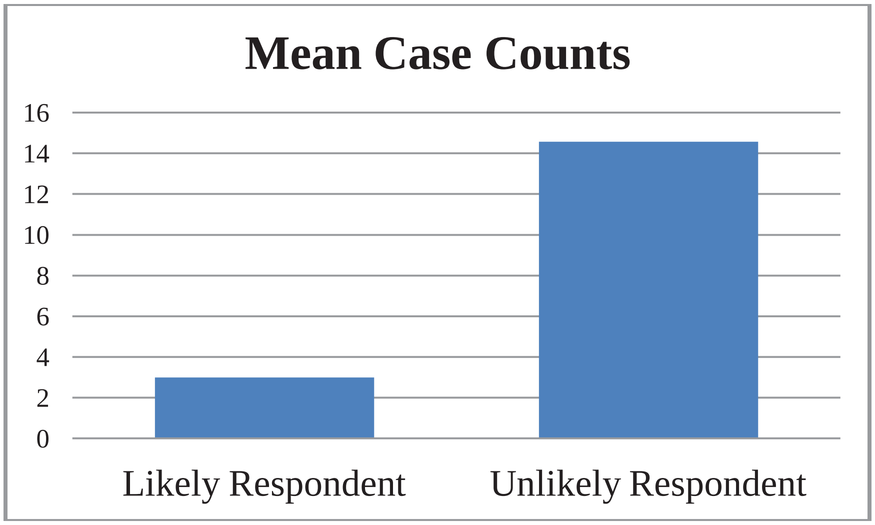

**Chart 1:** Bar chart of mean case counts for predicted likely and unlikely respondents.

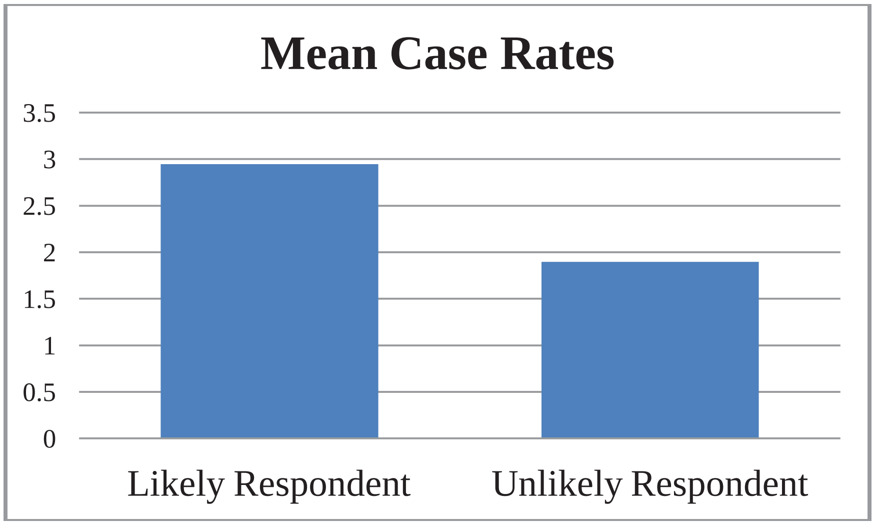

**Chart 2:** Bar chart of mean case rates for predicted likely and unlikely respondents.

## 3. Conclusion

Even though there is indication of nonresponse bias within the SOII, we are confident that it is most likely confined to the limits of this study, that is, confined to state and local government units within those states where reporting to the survey is voluntary. In states that require that the survey be submitted, BLS has response rates around 90%. Again, in 2011, the portion of states and governments represented is only 1.3% of the establishments and 6.2% of the employment. Because these percentages are so low, it is unlikely that nonresponse biases associated with state and local governments had a large impact on the national estimates for all industries. But for future work, public sector data for all states should be examined, as well as data for the private sector for the entire country.

***Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.***

## Appendix A: Maximum Likelihood Estimates and Standard Errors for Independent Variables in the Logistic Regression Model

| Variable | First Value | Second Value | Estimate | Standard Error |
|---|---|---|---:|---:|
| Intercept | | | -1.38801 | 0.24688 |
| Ownership | 20 | | -0.43731 | 0.03569 |
| Region | ATL | | 0.10874 | 0.27839 |
| Region | BOS | | -0.44317 | 0.37838 |
| Region | CHI | | 0.02286 | 0.28999 |
| Region | DAL | | 0.37211 | 0.26031 |
| Region | NSG | | 0.26776 | 0.25696 |
| Size Code | 1 | | 1.41464 | 0.07353 |
| Size Code | 2 | | 1.30677 | 0.07068 |
| Size Code | 3 | | 1.03580 | 0.06630 |
| Size Code | 4 | | 0.86676 | 0.06630 |
| Supersector | CON | | -0.63539 | 0.57795 |
| Supersector | EHS | | 1.11866 | 0.26890 |
| Supersector | FIA | | 1.85596 | 0.71363 |
| Supersector | INF | | 1.31867 | 0.59677 |
| Supersector | LEH | | -0.10780 | 0.66199 |
| Supersector | MFG | | 0.61265 | 0.47613 |
| Supersector | NRM | | 0.38391 | 1.04974 |
| Supersector | OTS | | -1.35575 | 1.11223 |
| Supersector | PAD | | 0.68336 | 0.26065 |
| Supersector | PBS | | 0.81594 | 0.53747 |
| Certainty | N | | -0.27588 | 0.04647 |
| Region * Supersector | ATL | CON | -0.98550 | 0.86109 |
| Region * Supersector | ATL | EHS | -1.39065 | 0.30801 |
| Region * Supersector | ATL | FIA | -0.59824 | 0.81714 |
| Region * Supersector | ATL | INF | -0.61359 | 0.65600 |
| Region * Supersector | ATL | LEH | 0.06490 | 0.71685 |
| Region * Supersector | ATL | MFG | -13.73234 | 373.26794 |
| Region * Supersector | ATL | NRM | 0.16217 | 1.12846 |
| Region * Supersector | ATL | OTS | 0.91361 | 1.15008 |
| Region * Supersector | ATL | PAD | -0.82713 | 0.30143 |
| Region * Supersector | ATL | PBS | -1.38059 | 0.62929 |
| Region * Supersector | BOS | CON | -1.75554 | 1.20686 |
| Region * Supersector | BOS | EHS | -0.39137 | 0.41586 |
| Region * Supersector | BOS | FIA | -0.46668 | 0.93673 |
| Region * Supersector | BOS | INF | -1.15239 | 0.80963 |
| Region * Supersector | BOS | LEH | 0.66492 | 0.85981 |
| Region * Supersector | BOS | MFG | 0.00000 | |

| | | | | |
|---|---|---|---|---|
| Region * Supersector | BOS | NRM | 0.00000 | |
| Region * Supersector | BOS | OTS | 1.33900 | 1.29738 |
| Region * Supersector | BOS | PAD | -0.05735 | 0.41322 |
| Region * Supersector | BOS | PBS | -0.57728 | 0.74599 |
| Region * Supersector | CHI | CON | 1.78212 | 0.74713 |
| Region * Supersector | CHI | EHS | -0.25737 | 0.31923 |
| Region * Supersector | CHI | FIA | -1.11196 | 0.81038 |
| Region * Supersector | CHI | INF | -0.90830 | 0.69311 |
| Region * Supersector | CHI | LEH | 0.92516 | 0.72141 |
| Region * Supersector | CHI | MFG | 13.07326 | 306.57790 |
| Region * Supersector | CHI | NRM | 12.80997 | 163.93051 |
| Region * Supersector | CHI | OTS | 2.92966 | 1.20302 |
| Region * Supersector | CHI | PAD | -0.09148 | 0.31259 |
| Region * Supersector | CHI | PBS | -0.46938 | 0.65497 |
| Region * Supersector | DAL | CON | 1.10092 | 0.59998 |
| Region * Supersector | DAL | EHS | -0.81931 | 0.28735 |
| Region * Supersector | DAL | FIA | -1.19754 | 0.74929 |
| Region * Supersector | DAL | INF | -1.04928 | 0.62448 |
| Region * Supersector | DAL | LEH | 0.19844 | 0.68209 |
| Region * Supersector | DAL | MFG | 1.32895 | 1.15603 |
| Region * Supersector | DAL | NRM | 2.66523 | 1.13344 |
| Region * Supersector | DAL | OTS | 1.78023 | 1.13618 |
| Region * Supersector | DAL | PAD | -0.37570 | 0.28056 |
| Region * Supersector | DAL | PBS | -0.63510 | 0.57005 |
| Region * Supersector | NSG | CON | 0.01403 | 0.60274 |
| Region * Supersector | NSG | EHS | -1.30526 | 0.28634 |
| Region * Supersector | NSG | FIA | -0.60185 | 0.74251 |
| Region * Supersector | NSG | INF | -0.66640 | 0.61969 |
| Region * Supersector | NSG | LEH | -0.10222 | 0.68015 |
| Region * Supersector | NSG | MFG | 0.00000 | |
| Region * Supersector | NSG | NRM | 1.56766 | 1.08459 |
| Region * Supersector | NSG | OTS | 1.24903 | 1.12746 |
| Region * Supersector | NSG | PAD | -0.96368 | 0.27889 |
| Region * Supersector | NSG | PBS | -1.09838 | 0.56121 |

## References

*BLS Handbook of Methods*, Chapter 9, September 2008, http://www.bls.gov/opub/hom/pdf/homch9.pdf

Huband, Erin. *Survey of Occupational Injuries and Illnesses Response Rate Analysis*, March 2010.

Selby, Philip; Burdette, Terry; Huband, Erin, *Overview of the Survey of Occupational Injuries and Illnesses Sample Design and Estimation Methodology*, October 2008.