Two-Phase Income Synthesis Using IPUMS Data

Henrik Olsson and Kevin Ros

May 2020

1 Abstract

The purpose of this study was to apply Bayesian methods to synthesize levels of income, with respect to an individual's socioeconomic, ethnic, and health characteristics. This paper focuses on the utility and risk evaluation methods applied to a publicly released dataset from 2018 annual National Health Interview survey, conducted by the United States Census. Variables age, sex, race, education level, hours worked, health insurance coverage, hours of sleep, and frequency of worry proved to be effective predictors collectively. The synthetic data approach involved a two-phase synthetic method, including linear and logistic regression models. Findings showed that...

2 Introduction

It is impossible to overstate the importance of data in today's world. Nearly every decision made by corporations and governments is based off conclusions drawn from data, in one way or another. However, many data sets structure information on an individual level (microdata) which leads to the possibility of disclosure risk (identifying attributes of an individual or identifying the individual). Not only does this inconvenience the individual but it may also lead to the disclosure of legally protected information, such as medical records. This is commonly avoided by using Bayesian methods to synthesize sensitive variables to reduce identification and attribute risks while maintaining the utility and relations between variables in the data set.

One particularly sensitive variable is income due to its uniqueness and potential for outliers. With certain data sets, income will contain a significant number of zeros (individuals with no income) along with non-zeros (individuals with income), which results a large spread between values. This can cause certain Bayesian synthesis models to lose effectiveness, thus yielding synthesized data with low utility. To prevent this, we attempt a two-step income synthesis process.

For the first step, we synthesize income categorically using logistic regression, where 0 indicates no income, and 1 indicates non-zero income. Then, in the second step, we synthesize all the non-zero income values using linear regression and the original income data. Combining the synthetic 0s from the first step and the synthetic income values from the second step yield a completely synthesized income.

This paper is outlined as follows. In the next section, we discuss the IPUMS data set used to test the two-step income synthesis method, as well as explain why the data set was chosen. Then, in Section 4, we detail the pre-processing of the data set and formally explain the two-step income synthesis method. We present the results of the synthesis, namely various utility and risk evaluation scores, in Section 5. Finally, we conclude with Section 6.

3 Background and Significance of the Research

72,832 samples were collected from the United States Census. Specifically, the data for this study was drawn from the National Health Interview Survey of 2018. The IPUMS Health Surveys harmonize these data and provide extensive information on the demographic, socioeconomic, and health experiences of individuals living in the U.S. The data were self-reported by random participants representing the U.S. population. Although the IPUMS provides surveys to the public every year, only the 2018 dataset was used for the study.

The dataset included a total of eight harmonized variables, including age, sex, race, education level, hours worked, health insurance coverage, hours of sleep, and frequency of worry. The dependent variable, income, was measured at the binary and nominal level. Specifically in the two-phase synthetic model, income of zeros and non-zeroes were used in the logistic regression, and nominal levels of income were used in the linear regression. Age, hours worked, hours of sleep and income (linear regression) represented continuous variables, while sex, race, education level, health insurance coverage, frequency of worry, and income (logistic regression) represented categorical variables. The variables were all chosen based on exploratory analyses of most sensitive variables to an individual's income. The analyses for this research were restricted to samples that had all survey fields answered fully with no missing values to be a part of the sample size.

The income variable was chosen to be synthesized due to the high sensitivity and potential disclosure risk. When disclosing sensitive information such as income, there is high risk that an intruder will be able to derive the confidential information given their knowledge of other characteristics. Through exploratory analyses, income was deemed the most sensitive among the nine variables. Specifically, the relationship between income and hours worked, as well as income and education levels were the most important relationships to preserve. With the addition of six more health and socioeconomic variables, the identification disclosure risk rises. Thus, the variable income was synthesized in order to protect the individual's privacy.

In order to protect the privacy of both individuals that receive and do not receive income, a two-phase synthesis measure was implemented. First, income is synthesized as binary values. This process is important to protect the privacy of unemployed individuals or students who receive zero income. Next, combining the synthetic 0s from the logistic regression to the linear regression with nominal non-zero income completes the full privacy protection method. Thus, the two-step method can be applied to various datasets with skewed values in order to maintain high utility, and lower disclosure risk.

4 Methods Used to Obtain and Analyze IPUMS Data

4.1 Data Preprocessing

For Education, Hours Worked, and Hours of Sleep, all rows that contained a variable value of 97 (Refused), 98 (Unknown- not ascertained), and 99 (Unknown - don't know) were removed. Similarly, this was done with Health Insurance Coverage and Frequency of Worry for values of 7, 8, and 9, as well as for Race but with values 970, 980, and 990. Additionally, all rows with NA values or 00 (Not In Universe) were also removed. This reduced the data set size from 72,832 entries to 14,287 entries. Finally, Race, Education, Health Insurance Coverage, and Frequency of Worry were re-coded to the values described in Figure 1 in the Appendix.

4.2 CatIncome Synthesis

-create dummy column for income (0s and 1s)
    -use jags / logistic regression to synthesize new 0s and 1s
    -call this CatIncome (categorical income)

4.3 Income Synthesis

-remove all 0s from original income
    -log it
    -use jags / linear regression to generate posterior draws
    -replace all non-zero income values in CatIncome with synthesized income value
    -results in synthetic income

5 Results of Analysis

Due Tuesday 4/28
    tables, charts, graphs, significance, confidence intervals, descriptive text

6 Discussion

Due Tuesday 4/28

A discussion of the research, the limitations of the current research, reasonableness of any assumptions made, possibilities of future work/studies that should be conducted, etc.

7 Appendix

Variable (Code name)	Description	Type	Value (*Re-coded value)	Synthesized
Income (EARNIMP1)	Total earnings from previous calendar year	Continuous	1 - 149,000	Yes
Age (AGE)	Age at time of survey	Continuous	18 - 85	No
Sex (SEX)	Participant sex	Categorical	1 = Male 2 = Female	No
Race (RACE)	Main racial background	Categorical	*1 = White 2 = Black/African American 3 = American Indian 4 = Asian 5 = Other races 6 = Two or more races	No
Education (EDUC)	Educational attainment	Categorical	*1 = 4 years of high school or less 2 = 1 - 4 years of college 3 = 5+ years of college	No
Hours Worked (HOURSWRK)	Total hours worked last week or usually	Continuous	1 - 95+	No
Health Insurance Coverage (HEALTH)	Health Insurance coverage status	Categorical	*1 = No, has coverage 2 = Yes, has no coverage	No
Hours of Sleep (HRSLEEP)	Usual hours of sleep per day	Continuous	0 - 24	No
Frequency of Worry (WORRY)	How often feel worried, nervous, or anxious	Categorical	*1 = Daily 2 = Weekly 3 = Monthly 4 = A few times a year 5 = Never	No

Figure 1: Variables used in the study. Data taken from 2018 IPUMS public use samples