NHIS Synthesis

Sarah Boese

April 28, 2020

Abstract

In this analysis, I look at a small subset of the National Health Interview Survey for 2018 pertaining to chronic conditions, worker status and missed work days. I use demographic data as unsynthesized predictors for this data. In so doing, I assume that demographics correlate person's probability of having a chronic condition. The main goal is to make sure that individuals who have chronic conditions are not made vulnerable to employers or healthcare insurance providers because of their participation in this survey.

1 Introduction

1.1 Research Questions

The National Health Interview Survey (NHIS) is a yearly survey conducted by the US Census Bureau concerning a broad range of health topics. My research has focused on the Sample Adult 2018 file, just one set of microdata released each year. This particular dataset includes 742 variables, many of which are sensitive. For my purposes, however, I am limiting my scope to only six variables: SEX (gender), RACERPI2 (race), AGE_P (age), FLA1AR (functional limitation), DOINGLWA (employment status) and WKDAYR (lost work days). I wish to answer the following questions: Are these demographic variables good predictors for FLA1AR and DOINGLWA? How can we best preserve the relationship between categorical variables describing functional limitation and work status to the continuous variable recording number of lost work days in a year? Moreover, does the entire synthesis process preserve to a satisfactory degree the "usefulness" of the data in further analysis? After I develop these models, the question becomes if such models still satisfactorily minimize risk of both identification and attribute disclosure?

1.2 Background/Significance of Research

Under Title 13, it is illegal for the US Census Bureau to "disclose or publish any census or survey information that identifies an individual or business". To that end, in order to release data that undergoes some conversion to ensure privacy. One way to protect participant's privacy is to release summary level data. However, releasing such data requires the addition of noise or perturbation to the data in order to ensure privacy. This is the focus of research in the field of differential privacy **cite here**.

However, from a modeling perspective, summary level data can be challenging to implement. Synthetic microdata is another research area looking to maintain privacy guarantees while also increasing utility for end users. *Microdata*, also called record-level or respondent-level data, is a collection of individual or business respondent data for a set list of variables/attributes. In this paper, we will use Bayesian modeling to generate synthetic data for variables we find at risk for disclosing sensitive information about individuals who participated in the 2018 National Health Interview Survey.

First, we must formalize what we mean when we say disclosure risk. *Disclosure Risk* is the risk posed by an intruder or attacker who uses a publicly available database to derive confidential information about

individuals in the database **cite Hu here**. We will consider two types of disclosure risks: identification and attribute disclosure. *Identification Disclosure* is an intruder using mircrodata to find out about the identity of an individual that the intruder is specifically looking for. We say this is like your neighbor finding out information about you by identifying you in a dataset. *Attribute Disclosure*, by contrast, consists of an intruder correctly inferring the true value of one or more unknown variable(s) in an individual's response. We will use measures to measure these risks before and after synthesis to see if we have introduced privacy protection to the synthetic micro-level data.

From a statistical standpoint, it is important that we release usable data. That is, it is pertinent that we maintain those relationships between different response variables that would be important for users who use the data for analysis.

2 Methods Used to Obtain Data

As mentioned above, I obtained the variables of interest from the NHIS Sample Adult data set for 2018. Below I include descriptions of both the predictor and to-be-synthesized variables:

2.1 Predictor Variables

Predictor Variable Table	
Variable	Information
SEX	1 = male, 2 = female
RACEPI2	1 = white only, 2 = Black, 3 = AIAN only, 4 = Asian only, 5 Race group not releasable, 6 = Multiple Race
AGE_P	18-84 = 18-84 years, 85 = 85+ years

The **SEX** variable is a binary variable used to represent gender. The NHIS does not capture the multiplicity of gender identities that members of society can have, instead it is restricted to the gender binary. Because we are using data collected by the government, we will have to conform to their restrictions.

The **RACERPI2** variable is a six-level categorical variable describing race. It is not the only variable pertaining to race in the Sample Adult dataset, however, it is the most up-to-date in terms of OMB standards.

Finally, **AGE_P** is the age variable in the survey. It is integer valued and top coded at 85 years:

2.2 Synthesized Variables

The following three variables, (FLA1AR, DOINGLWA and WKDAYR) I intend to synthesize during my analysis. All of these variables, like the above, have non-recorded options represented as variables. I will not be considering data points that have such unrecorded options.

Synthesized Variable Table	
Variable	Information
FLA1AR	1 = limited in any way, 2 = not limited in any way
DOINGLWA	1 = Working for pay at a job or business, 2 = With a job or business but not at work, 3 = Looking for work, 4 = Working, but not for pay, at a family-owned job or business, 5 = Not working and not looking for work
WKDAYR	000 = None, 001-366 = 1-366 days

FLA1AR is a binary varibale denoting if a participant has any functional limitation. We chose not to look at whether or not that limitation was chronic since over 95 percent of people experiencing a functional

limitation has one such limitation which is chronic. As part of the cleaning process, I delete the rows such where $FLA1AR = 3$.

In the NHIS Sample adult dataset, the **DOINGLWA** is a eight-level categorical variable trying to capture a participants work status in the week before participating in this study. Three of those options are non-recorded, so they are not considered.

Finally, we are interested in the number of lost work days participants needed to take during the past year due to health problems. Only 1.7% of participants in this survey loose more than 50 workdays a year and only 0.6% loose more than 100 work days. Participants under these constraints are particularly vulnerable to being identified as members of this survey. The **WKDAYR** variable takes the following values:

3 Analysis

3.1 Models

We use three distinct types of models for our synthesized variables depending on the type of outcome each variable has. For the binary variable **FLA1AR** we use a Bayesian logistic regression with SEX and RACERPI2 as predictors. This model is appropriate as it fits binary outcome and allows us to use categorical predictor variables. We outline the model below:

$$\begin{aligned} Y \sim &\text{Bernoulli}(p) \\ \text{logit}(p) = &\beta_1 + \beta_2 \cdot SEX_{female} + \beta_3 \cdot RACE_{black} + \beta_4 \cdot RACE_{AIAN} + \beta_5 \cdot RACE_{asian} \\ &+ \beta_6 \cdot RACE_{not\ releasable} + \beta_7 \cdot RACE_{multiple} \\ \beta_i \sim &\text{Normal}(\mu, \sigma) \end{aligned}$$

In the model we implemented, we let $\mu = 0$ and $\sigma = 0.1$.

We use a similar approach for modeling the **DOINGLWA** variable. Instead of using simple logistic regression, we must use multinomial logistic regression as DOINGLWA has five possible outcomes. This way we can fit all outcomes and give them predictor variables:

$$Y \sim \text{Categorical}(p)$$

For the count of missed work days in a year, ie. the **WKDAYR** variable, we will use Poisson Regression. Similar to the logistic regressions we ran on the previous variables, we can also give that Poisson regression linear predictors. However, we will use only the two to-be-synthesized variables as predictor variables instead of the demographic variables.

$$\begin{aligned} Y \sim &\text{Poisson}(\lambda) \\ \log(\lambda) = &\beta_1 + \beta_2 \cdot FLA1AR_{no\ lim} + \beta_3 \cdot DOINGLWA_{working} + \beta_4 \cdot DOINGLWA_{vacation} \\ &+ \beta_5 \cdot DOINGLWA_{looking} + \beta_6 \cdot DOINGLWA_{work\ not\ for\ pay} + DOINGLWA_{not\ looking} \end{aligned}$$

3.2 Sequential Synthesis

To create our synthetic dataset, we use a method similar to that used to preform posterior predictive checks. That is, we take one draw of the posterior parameters and we use the distribution defined by those parameters to generate the synthetic day. For example, in the FLA1AR synthesis model, we consider one set of $\{\beta_1, \beta_2, \ldots, \beta_7\}$ to calculate p_i for each row of predictors. We then pull our synthetic value $\tilde{Y}$ from the

Bernoulli distribution defined by p_i.

In order to maintain the relationship between our synthesized variables, we use some of them as predictors. This means that we first fit our model with the un-synthesized, or original predictor variables. Then we generate our synthetic data in a sequential manner (where * denotes synthesized variables):

$$\pi(FLA1AR^* \mid SEX, RACERPI2)$$
$$\pi(DOINGLWA^* \mid FLA1AR^*, SEX, RACEPI2)$$
$$\pi(WKDAYR^* \mid FLA1AR^*, DOINGLWA^*).$$

We call this process *Sequential Synthesis*. It is very similar to the process of multiple imputation for missing data.

4 Results

In general, these modeling methods result in a synthetic data set that has quite high utility. Visually, we see that the logistic and multinomial logistic regressions worked very well to generate a similar distribution of synthetic data to the original. However, the Poisson regression did not capture the skewed WKDAYR data, though it did maintain high utility given the utility measures we will discuss in the following section. First, let's consider the synthesized FLA1AR.

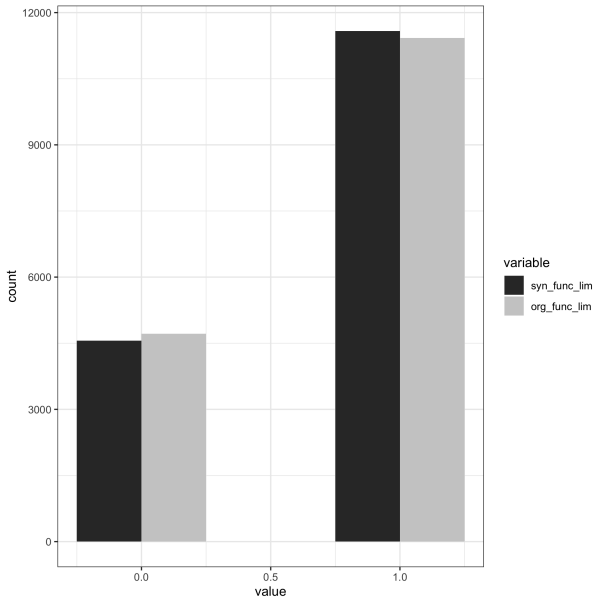

Figure 1: FLA1AR Synthesis

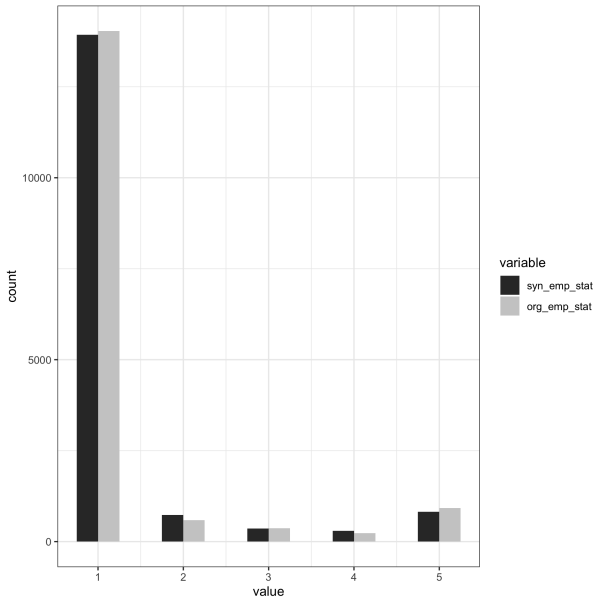

Figure 2: DOINGLWA Synthesis

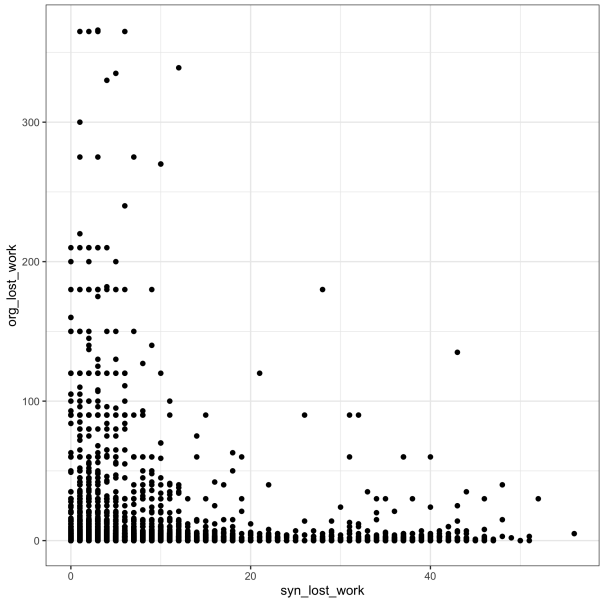

Figure 3: WKDAYR Synthesis