Methods for Utility Evalation part 1

Jingchen (Monika) Hu

Vassar College

Statistical Data Privacy

Outline

1. Introduction
2. Global utility measure #1: propensity scores (*pMSE*)
3. Global utility measure #2: empirical cumulative distribution function (eCDF)
4. Summary and References

Outline

1. Introduction
2. Global utility measure #1: propensity scores (*pMSE*)
3. Global utility measure #2: empirical cumulative distribution function (eCDF)
4. Summary and References

Recap

- Lecture 1 Overview of synthetic data: various aspects of creating synthetic data for microdata privacy protection

Recap

- Lecture 1 Overview of synthetic data: various aspects of creating synthetic data for microdata privacy protection
- Lectures 2 & 3 Introduction to Bayesian modeling: nuts and bolts of Bayesian modeling
  - The foundation of Bayesian inference: prior, likelihood, Bayes' rule (discrete and continuous), and posterior
  - Markov chain Monte Carlo (MCMC): estimation and diagnostics
  - Posterior predictive and synthetic data
  - Case study: gamma-Poisson conjugate model, Poisson regression model, posterior prediction, prior choices, posterior inference, MCMC, MCMC diagnostics, using the `brms` package

Recap

- Lecture 4 & 5 Bayesian synthesis models
  - Synthesizing continuous variables: Bayesian simple / multiple linear regressions
  - Synthesizing binary variables: Bayesian logistic regression
  - Synthesizing categorical variables: Bayesian multinomial logistic regression
  - Synthesizing count variables: Bayesian Poisson regression
  - In all cases
    - ★ Prior (default priors in `brms`)
    - ★ MCMC estimation and diagnostics
    - ★ Posterior predictions for synthetic data generation (writing R functions); sequential synthesis
    - ★ Simple utility checks
  - Synthesizing multivariate categorical variables: the DPMPM and the `NPBayesImputeCat` R package

The CE sample

- Our sample is from the 1st quarter of 2019, containing five variables on 5133 CUs

Variable Name	Variable information
UrbanRural	Binary; the urban / rural status of CU: 1 = Urban, 2 = Rural.
Income	Continuous; the amount of CU income before taxes in past 12 months (in *USD*).
Race	Categorical; the race category of the reference person: 1 = White, 2 = Black, 3 = Native American, 4 = Asian, 5 = Pacific Islander, 6 = Multi-race.
Expenditure	Continuous; CU's total expenditures in last quarter (in *USD*).
KidsCount	Count; the number of CU members under age 16.

Plan for this lecture

- Two general types of utility: global and analysis-specific (Lecture 1)

1. Global utility: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
2. Analysis-specific utility: Evaluate whether synthetic data users can obtain statistical inferences on the synthetic data that are similar to those obtained on the confidential data

Plan for this lecture

- Two general types of utility: global and analysis-specific (Lecture 1)

1. Global utility: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
2. Analysis-specific utility: Evaluate whether synthetic data users can obtain statistical inferences on the synthetic data that are similar to those obtained on the confidential data

- In this lecture, we focus on global utility evaluation methods, with illustrations to the synthetic CE from Lectures 4 & 5

Discussion question: Given your readings and previous lectures, what are global utility evaluation methods you have seen?

Overview

- Global utility measures focus on identifying the distributional differences between the confidential data and the synthetic data

Overview

- Global utility measures focus on identifying the distributional differences between the confidential data and the synthetic data
- The key to all these global measures is in discriminating between the confidential and the synthetic data using common statistical tools

Outline

1. Introduction
2. Global utility measure #1: propensity scores (*pMSE*)
3. Global utility measure #2: empirical cumulative distribution function (eCDF)
4. Summary and References

Propensity scores

- **Propensity scores** measure the probability for individuals in a dataset being assigned to a specific treatment group given their information on other variables
- Consider two groups: group A receiving a treatment and group B not receiving it
- Given other measured covariates, we can estimate for each individual the probability that they were assigned to group A

Propensity scores

- **Propensity scores** measure the probability for individuals in a dataset being assigned to a specific treatment group given their information on other variables
- Consider two groups: group A receiving a treatment and group B not receiving it
- Given other measured covariates, we can estimate for each individual the probability that they were assigned to group A
- If the probability distributions differ for the two groups, this indicates that the individuals in the two groups differ with respect to these covariates

Propensity scores matching vs synthetic data

- Propensity scores are used in **propensity score matching**, a technique commonly employed in causal inference aimed at reducing bias due to confounding variables when estimating the effect of a treatment, policy, or other interventions in an observational study

Propensity scores matching vs synthetic data

- Propensity scores are used in **propensity score matching**, a technique commonly employed in causal inference aimed at reducing bias due to confounding variables when estimating the effect of a treatment, policy, or other interventions in an observational study
- To evaluate the quality of synthetic datasets, we use them to investigate whether the synthetic observations significantly differ from the original observations
- We follow Woo et al. (2009) and Snoke et al. (2018) and introduce the *pMSE* as the propensity scores utility measure

pMSE calculation

- When used as a utility measure for synthetic data, the treatment of interest is that the observation is part of the generated synthetic dataset

pMSE calculation

- When used as a utility measure for synthetic data, the treatment of interest is that the observation is part of the generated synthetic dataset

- The steps of computing *pMSE* metric and using it for evaluation of global utility:

1. Merge the confidential and the synthetic datasets. Assume the confidential dataset has n_c records while the synthetic dataset has n_s records. We merge the two datasets by stacking them together, resulting in a merged dataset of dimension $(n_c + n_s)$-by-r, where r is the number of variables
2. Add an additional variable, S. For record i $(i = 1, \cdots, (n_c + n_s))$, set $S_i = 0$ if it comes from the confidential dataset and $S_i = 1$ if it comes from the synthetic dataset

pMSE calculation

3. For each record i $(i = 1, \cdots, (n_c + n_s))$, estimate the probability of it being in the synthetic dataset by fitting a model using available predictors in the datasets. This probability is the estimated propensity score, denoted as $\hat{p}_i$.

pMSE calculation

4. Compare the distributions of the propensity scores in the confidential and the synthetic datasets by computing the propensity score mean-squared error, known as *pMSE*, which is the mean-squared difference between the estimated propensity probabilities and the probability of a record being synthetic if original and synthetic observations were interchangeable, as:

$$pMSE = \frac{1}{n_c + n_s} \sum_{i=1}^{n_c+n_s} (\hat{p}_i - c)^2, \tag{1}$$

where $(n_c + n_s)$ is the number of records in the merged dataset, $\hat{p}_i$ is the estimated propensity score for unit i, and c is the proportion of units with synthetic data in the merged dataset, $c = n_s/(n_c + n_s)$

pMSE calculation

- When $n_c = n_s$, $c = 1/2$
- When multiple synthetic datasets are generated, i.e., $m > 1$, *pMSE* would be calculated for each synthetic dataset and we can report the average *pMSE* value as the overall propensity score utility metric

pMSE implications

$$pMSE = \frac{1}{n_c + n_s} \sum_{i=1}^{n_c+n_s} (\hat{p}_i - c)^2 \tag{2}$$

- For a synthetic dataset of high utility, the classification model used in Step 3 will not be able to distinguish between the confidential and the synthetic datasets

- In such cases, the propensity scores will all be close to 0.5, leading to a *pMSE* of zero

Discussion question: Does a high *pMSE* or a low *pMSE* indicate high utility? Why?

pMSE implications

- Larger values of the *pMSE* indicate a lower level of utility
- The largest possible value of *pMSE* when $n_c = n_s$ is $1/4$, when each $\hat{p}_i$ is equal to 0 or 1

Additional comments regarding *pMSE*

- We can use any classification model in Step 3, for example, logistic regression, classification tree, random forest, support vector machines (SVM)

- As pointed out by Snoke et al. (2018), if a logistic regression model is used, we can calculate the null distribution of the *pMSE*, which allows us to construct a hypothesis test (check out Nowok et al. (2020) for details)

Additional comments regarding *pMSE*

- The choice of model for classification will have a big impact of the calculated *pMSE*
  - Using a poor model will make it difficult to distinguish between the original and the synthetic dataset, artificially indicating good utility for the synthetic dataset

Additional comments regarding *pMSE*

- The choice of model for classification will have a big impact of the calculated *pMSE*
  - Using a poor model will make it difficult to distinguish between the original and the synthetic dataset, artificially indicating good utility for the synthetic dataset
- In practice, the model used for classification should include at least all of the variables which were synthesized

Additional comments regarding *pMSE*

- The choice of model for classification will have a big impact of the calculated *pMSE*
  - Using a poor model will make it difficult to distinguish between the original and the synthetic dataset, artificially indicating good utility for the synthetic dataset
- In practice, the model used for classification should include at least all of the variables which were synthesized
- Moreover, it is recommended to include second-order and even third-order interactions between the variables in the model if we want to test whether the relationships in the synthetic version of the confidential dataset are preserved (Woo et al. (2009), Snoke et al. (2018))

pMSE calculation for synthetic CE Expenditure

- We evaluate the *pMSE* global utility of synthetic CE expenditure example from Lecture 4 using a logistic regression with the main effects and first-order interactions of all 5 variables as a classification model in the CE sample
- See the unshown code for creating the synthetic dataset `SLR_synthetic_one_final` (synthesized `Expenditure` exponentiated from synthesized `LogExpenditure`)
- Important note: typically we evaluate utility of synthetic data of the original scales

```
SLR_synthetic_one_final[1:3, ]
```

```
## # A tibble: 3 x 7
##   UrbanRural Income  Race Expenditure KidsCount LogIncome LogExpenditure
##        <dbl>  <dbl> <dbl>       <dbl>     <dbl>     <dbl>          <dbl>
## 1          1  73720     1      12831.         3     11.2           10.2
## 2          1  12000     1       2327.         2      9.39           8.91
## 3          1  20000     1       2975.         0      9.90           9.06
```

pMSE calculation: merge and add *S* variable

```
merged_data <- rbind(CEdata,
                     SLR_synthetic_one_final)

merged_data$S <- c(rep(0, n), rep(1, n))
```

pMSE calculation: compute estimated propensity scores with a logistic regression

- Main effects and first-order interactions as predictors
- Use `glm()` function from the `stats` package with `family = "binomial"` for a logistic regression (non-Bayesian)

```
log_reg <- stats::glm(formula = S ~ (as.factor(UrbanRural) + Income +
                        as.factor(Race) + Expenditure + KidsCount)^2,
                      family = "binomial",
                      data = merged_data)
```

pMSE calculation: compute estimated propensity scores with a logistic regression

- Use the `predict()` function to calculate the estimated $\hat{p}_i$ for all records in the merged dataset

```
pred <- stats::predict(log_reg,
                       data = merged_data)
probs <- exp(pred) / (1 + exp(pred))
```

pMSE calculation: calculate *pMSE*

```
pMSE <- 1 / (2 * n) * sum((probs - 1 / 2)^2)
pMSE
```

```
## [1] 0.001236854
```

- The calculated propensity score utility measure *pMSE* is small and close to 0, meaning that the logistic regression model cannot really distinguish between the confidential and the synthetic datasets, indicating a high level of utility of our simulated synthetic data

Discussion question: In Lecture 5 we visually checked synthetic `KidsCount` - do you think the calculated *pMSE* there would be higher or lower than here? Why?

Outline

1 Introduction

2 Global utility measure #1: propensity scores (*pMSE*)

3 Global utility measure #2: empirical cumulative distribution function (eCDF)

4 Summary and References

eCDF definition

- An empirical distribution function, also commonly known as an **empirical cumulative distribution function** (**empirical CDF**), is the distribution function associated with the empirical measure of a sample

- It is a discrete distribution function which considers every observation in the sample to be an equally likely outcome:

Let $(y_1, \cdots, y_n)$ be the n sample observations. The eCDF is defined as:

$$\hat{F}_n(t) = \frac{\text{the number of observations in the sample} \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{y_i \leq t}, \tag{3}$$

where $\mathbf{1}_{y_i \leq t}$ is the indicator of event $y_i \leq t$, and n the total sample size.

eCDF for synthetic data

- In the context of global utility of synthetic data, the empirical CDF measures rely on comparing the empirical distributions of the synthetic dataset with that of the confidential dataset

- Two similar samples should have similar empirical CDFs, and this can be used to obtain a global utility measure of synthetic data

eCDF calculation

- We follow Woo et al. (2009) and start with merging the two datasets (note that the merging step is not necessary to calculate the empirical CDFs)

1. Merge the confidential and the synthetic datasets, resulting a merged dataset of dimension $(n_c + n_s)$-by-r.
2. Estimate the empirical CDF distribution of the confidential dataset, denoted as $ecdf^O$, and that of the synthetic dataset, denoted by $ecdf^S$.

eCDF calculation

3. For record i $(i = 1, \cdots, (n_c + n_s))$ in the merged dataset, estimate its **percentile** under the empirical CDF distribution of the confidential dataset $ecdf^O$, denoted as p_i^O, and its percentile under the empirical CDF distribution of the synthetic dataset $ecdf^S$, denoted as p_i^S.

eCDF calculation

4. Use the following two measures:

$$U_m = \max_{1 \le i \le (n_c+n_s)} |p_i^O - p_i^S|, \tag{4}$$

$$U_a = \frac{1}{(n_c + n_s)} \sum_{i=1}^{(n_c+n_s)} (p_i^O - p_i^S)^2, \tag{5}$$

where U_m is the maximum absolute difference between the empirical CDFs, and U_a is the average squared differences between the empirical CDFs

eCDF implications

- The smaller the values of U_m and U_a, the higher the similarity level between the confidential and the synthetic data, indicating high utility
- The larger the values of U_m and U_a, the lower the similarity level between the confidential and the synthetic data, indicating low utility

eCDF calculation for synthetic CE Expenditure

- We evaluate the eCDF global utility of synthetic CE expenditure example from Lecture 4

eCDF calculation: estimate the two eCDFs

- Merge the two datasets

```
merged_data <- rbind(CEdata,
                     SLR_synthetic_one_final)

merged_data$S <- c(rep(0, n), rep(1, n))
```

- Use `ecdf()` function from the `stats` R package

```
ecdf_orig <- stats::ecdf(CEdata$Expenditure)
ecdf_syn <- stats::ecdf(SLR_synthetic_one_final$Expenditure)
```

eCDF calculation: estimate the percentiles in the merged dataset

```
percentile_orig <- ecdf_orig(merged_data$Expenditure)
percentile_syn <- ecdf_syn(merged_data$Expenditure)
```

eCDF calculation: calculate U_m and U_a

```
ecdf_diff <- percentile_orig - percentile_syn

Um <- max(abs(ecdf_diff))
Um
```

```
## [1] 0.02474187
```

```
Ua <- mean(ecdf_diff^2)
Ua
```

```
## [1] 0.0001698552
```

- The calculated empirical CDF utility measures U_m and U_a are small, meaning that the empirical CDFs of the confidential dataset and of the synthetic dataset are similar

Additional comments regarding eCDF

- The empirical CDF and this global utility evaluation framework can be generalized to more than one variable
- However, in practice, it is often challenging to estimate joint empirical CDFs for multiple variables (no widely-used functions / packages for this purpose, such as `ecdf()` from the `stats` package we used for the univariate case)

Outline

1 Introduction

2 Global utility measure #1: propensity scores (*pMSE*)

3 Global utility measure #2: empirical cumulative distribution function (eCDF)

4 Summary and References

Summary

- Global utility measures
  - Propensity scores and *pMSE*, calculation, and implications
  - eCDF, calculation, and implications
  - Applied to the synthetic Expenditure in CE sample

Summary

- Global utility measures
  - Propensity scores and *pMSE*, calculation, and implications
  - eCDF, calculation, and implications
  - Applied to the synthetic Expenditure in CE sample
- No homework! But you should start checking global utility of your simulated synthetic data for your project
- Lecture 7: Methods for utility evaluation part 2
  - Drechsler (2011), Section 7.1

References I

Drechsler, J. 2011. Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics 201, Springer.

Nowok, B., G. M. Raab, C. Dibben, J. Snoke, and C. van Lissa. 2020. Synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control.
https://cran.r-project.org/web/packages/synthpop/index.html.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." Journal of the Royal Statistical Society, Series A (Statistics in Society) 181 (3): 663–88.

Woo, M. J., J. P. Reiter, A. Oganian, and A. F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." The Journal of Privacy and Confidentiality 1 (1): 111–24.