Synthesis Method - Part 2

Synthetic Linear Regression Model

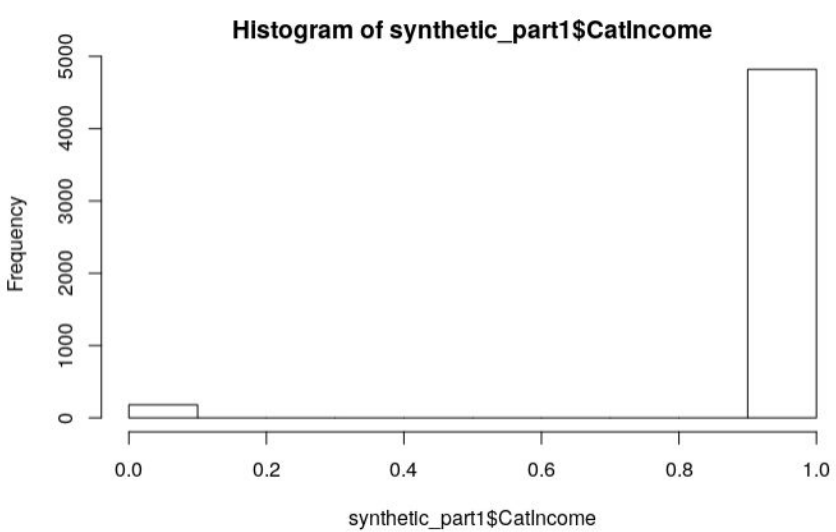
- Goal is to synthesize income using the data and CatIncomeSyn from Part 1
- First, copy/remove all non-zero values from OrigIncome, named OrigIncomeNoZero
- Then, use Synthetic Linear Regression model and OrigIncomeNoZero to synthesize OrigIncomeSyn (which has same dimensions as OrigIncome and CatIncomeSyn)
- Finally, modify OrigIncomeSyn as follows:
  - OrigIncomeSyn[i] = 0 if CatIncomeSyn[i] = 0
  - OrigIncomeSyn[i] = OrigIncomeSyn[i] if CatIncomeSyn[i] = 1
- OrigIncomeSyn now consists of the synthesized income values

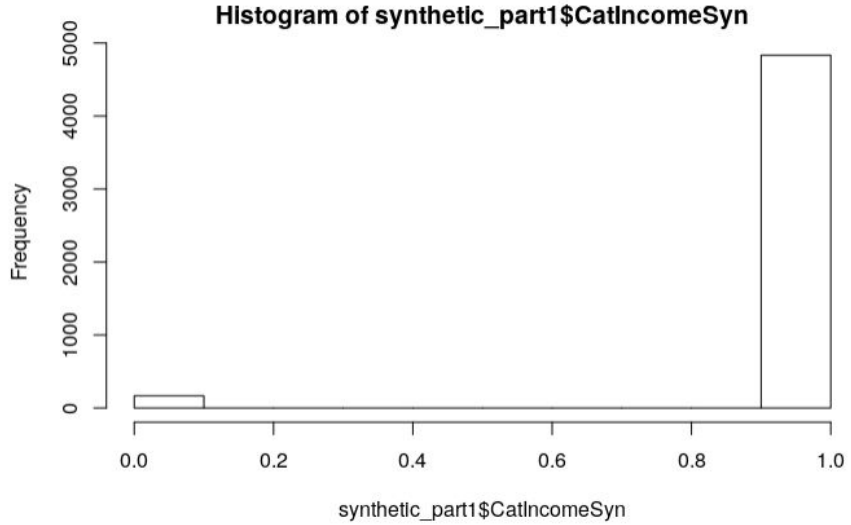
Variable Selection

Variable	Description	Type	Values	Synthesized
Income	Total earnings from previous calendar year	Categorical	0, 1	Yes
		Continuous	1 - 149,000	
Age	Age at time of survey	Continuous	18 - 85	No
Sex	Participant sex	Categorical	1 = Male 2 = Female	No
Race	Main racial background	Categorical	1 = White 2 = African American 3 = American Indian 4 = Asian 5 = Other races 6 = Two or more races	No
Education	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	Total hours worked last week or usually	Continuous	1 - 95+	No
Health insurance coverage	Health Insurance coverage status	Categorical	No, has coverage Yes, has no coverage	No
Hours of sleep	Usual hours of sleep per day	Continuous	0 - 24	No
Frequency of worry	How often feel worried, nervous, or anxious	Categorical	1 = Daily 2 = Weekly 3 = Monthly 4 = A few times a year 5 = Never	No

Part 1: Synthetic Logistic Regression

Original Income vs Synthetic Income Comparison





	0	1	Total
CatIncome	181	4819	5000
CatIncomeSyn	168	4832	5000

Part 2: Synthetic Linear Regression

Significance of the Research

We compare the utility and risk measures of the two-phase income synthesis process alongside those of the single-phase income synthesis process

Utility Evaluation - Global Measures

Single-phase income synthesis

Propensity score

- $U'_p = 0.000567386$

Cluster analysis

- $U'_c = 0$

Empirical CDF

- $U'_m = 0.24126$

  $U'_s = 0.01930874$

Two-phase income synthesis

Propensity score

- $U_p = 2.41566e - 05$

Cluster analysis

- $U_c = 0$

Empirical CDF

- $U_m = 0.10063$

  $U_s = 0.002671163$

All measures are averages from m = 20 synthetic datasets

Utility Evaluation - Global Measures

Single-phase income synthesis

Propensity score

- $U'_p = 0.000567386$

Two-phase income synthesis

Propensity score

- $U_p = 2.41566e-05$

The propensity score for the two-phase measure implies that p_i ≈ c across both the original and synthetic data, indicating high utility

The single-phase propensity score is slightly lower indicating lower utility

Utility Evaluation - Global Measures

Single-phase income synthesis

Cluster analysis

- $U_c' = 0$

Two-phase income synthesis

Cluster analysis

- $U_c = 0$

We set G, the number of clusters, equal to 50.

The data utility for both the single-phase and two-phase syntheses indicate high data utility

Utility Evaluation - Global Measures

Single-phase income synthesis

Empirical CDF

- $U'_m = 0.24126$

  $U'_s = 0.01930874$

Two-phase income synthesis

Empirical CDF

- $U_m = 0.10063$

  $U_s = 0.002671163$

U_m refers to the maximum absolute difference
U_s refers to the average squared difference

Although U_m and U_s are relatively close to 0 in both synthesis, their values indicate that our original and synthetic datasets have non-trivial differences between their distributions

The values for single-phase income synthesis is relatively higher, indicating a decrease in utility

Utility Evaluation - Analysis-specific Measures

Single-phase income synthesis

- $mean'_{syn} = 54297,$
  $median'_{syn} = 24570.07.$
- $interval'_{95} = [52640.95, 55953.05]$
- $I' = -0.5692832$

Given the 95% CI of [48941.63, 51137.53] for the original dataset, the intervals do not overlap resulting in a negative value
Indicates a decrease in utility

Two-phase income synthesis

- $mean_{syn} = 50537.89$
  $median_{syn} = 33954.89$
- $interval_{95} = [49237.86, 51837.91]$
- $I = 0.7978614,$

Indicates relatively high utility, although it seems to reflect the disparities in median between the original and synthetic datasets

Utility Evaluation - Analysis-specific Measures

Figure 1: Violin plot of original (0) and two-phase synthetic (1) income

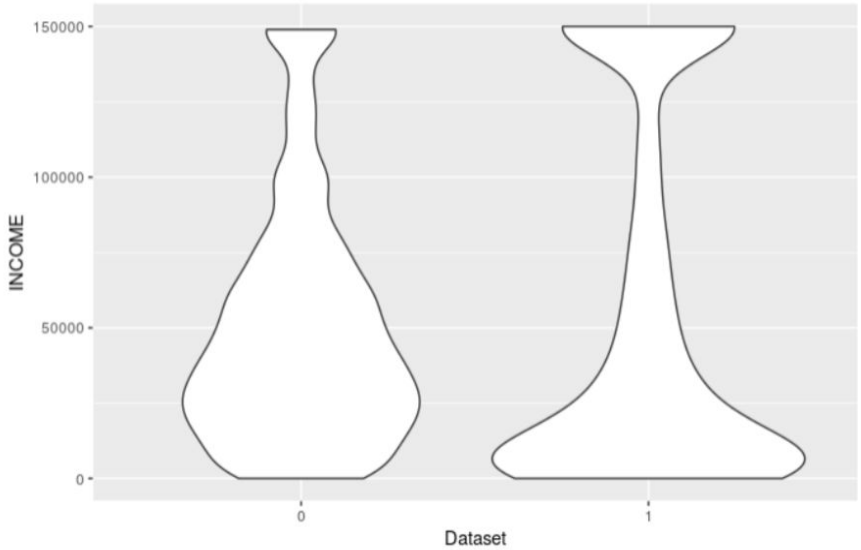


Figure 2: Violin plot of original (0) and single-phase synthetic (1) income

Identification Disclosure Risk Evaluation

Radius	Measure	Single-Phase	Two-Phase
0.1	Expected Match Risk	0.264441	0.4580699
	True Match Rate	0	2e-05
	False Match Rate	1	0.9992752
0.2	Expected Match Risk	0.3296083	0.3916344
	True Match Rate	0	1e-05
	False Match Rate	1	0.9993243
0.5	Expected Match Risk	0.315921	0.3566615
	True Match Rate	0	0
	False Match Rate	1	1
0.9	Expected Match Risk	0.323256	0.3510375
	True Match Rate	0	0
	False Match Rate	1	1

For each radius value, there is a significant expected match risk, but the risk is slightly lower for the single-phase income values.
Overall, the risk is slightly lower for single-phase income synthesis, as there is a lower correlation between the original and the synthesized income values.

Discussion - Limitations

- Assumed random missing for missing values and removed those observations.
  - Missing values may carry information about the observations themselves
  - Significantly decreased our sample size

- We chose the variables based on our own intuition due to possible correlations and sensitivity to the variable income
  - There may be additional variables that can be implemented to improve the utility evaluation and lower the risk measures
  - Variables related to medical care access, health behaviors, occupation, and family interrelationships can provide a more accurate model.

- No result for attribute disclosure risks

- Uncertainty on two-phase model application to other datasets
  - Applying to only partially synthetic data

Discussion - Future Research

- Random missing for missing values: Further exploration should be conducted by including more observations in the sample
- Variable selection: Develop measures to assess what variables hold the most sensitive relationship with the response variable
- Attribute disclosure risks
- Uncertainty on two-phase model application to other datasets: For fully synthetic data, we can use a Bayesian logistic regression to synthesize the binary income values, then implement a sequential synthesis for the second phase. Then we can sequentially synthesize each variable at a time, given the previously synthesized variable
- Top-coding: Implement in order to protect the privacy of an individual's income, specifically useful for datasets with outliers

References

[1] Jorg Drechsler and JP Reiter. Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey. Journal of Official Statistics, 25(4):589, 2009.

[2] Jingchen Hu. Bayesian estimation of attribute and identification disclosure risks in synthetic data. arXiv preprint arXiv:1804.02784, 2018.

[3] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality, 1(1), 2009.