Datasheet for '2019 Revised Supplemental Poverty Measure Research Dataset'*

Mary Cheng

April 17, 2024

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

   - The dataset was created to enable analysis of the United States household poverty status. The dataset analyzes poverty status using various demographic and financial factors. There is no specific gap that needed to be filled.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

   - United States Census Bureau

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

   - No direct funding was received for this project.

4. *Any other comments?*

   - N/A

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - Each row of the main dataset represents an individual in a household.

---

*Code and data are available at: https://github.com/marycx/us_poverty_analysis_2019.git

2. *How many instances are there in total (of each type, if appropriate)?*

   - There are 157,959 instances.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset is a sample. The larger set should be all the people in the United States of America. The sample is not fully representative of the larger set. Even though the sample takes various demographic variables into consideration, there are still people in the US who are not surveyed. This creates a potential problem that some people with different age, race, and background might not be accounted for.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of different demographic and financial information of each participant in this survey.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - No. There is unique identifier for each household, however, if there are more than one person in the household, these people will have the same identifier since they live in the same household.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - No.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - No, since there is no unique identifier associated with each individual instance.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- There is some human uncertainties in the dataset, for example, respondents might report false information in the survey.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - Self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes, age and race.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Maybe. If you include other information like government benefit programs with this information, you might be able to cross reference some individuals.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - Yes, financial information is contained in this dataset.

16. *Any other comments?*

    - No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

   - The data was mainly collected by surveys by United States Census Bureau. Portions of SPM resources, such as values of taxes and some non-cash benefits, were not asked of respondents and were estimated using models (Service 2022).

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

   - Surveys and a computer model.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - It is not a sample.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - workers from United States Census Bureau.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The data was created by the United States Census Bureau in 2019.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - No.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Not sure.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - Yes, it was done by survey (Service 2022).

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - Not sure.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - N/A

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No.

12. *Any other comments?*

    - N/A

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

   - No.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

   - N/A

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   - N/A

4. *Any other comments?*

   - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

   - Yes, for 2019 poverty status analysis by US government.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other acess point.*

   - No.

3. *What (other) tasks could the dataset be used for?*

   - The dataset can also be analyzed to show relationship between different data features and see how those relate to each other.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - There are some financial information in there that might affect individuals. For example, if the bank knows that you were in poverty for years, they are less likely to give you a loan.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - No.

6. *Any other comments?*

   - N/A

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - The dataset is available through United States Census Bureau official website.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- STATA / SAS data download.

3. *When will the dataset be distributed?*

   - It is available now.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - No.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

   - No.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - No.

7. *Any other comments?*

   - N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - No one.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - N/A

3. *Is there an erratum? If so, please provide a link or other access point.*

   - No.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - No.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - No.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - No.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - The dataset is on government website. You may contact the United States Census Bureau officials in ways listed on this website: https://www.census.gov/about/contact-us.html

8. *Any other comments?*

   - N/A

Reference

Service, Congressional Research. 2022. *The Supplemental Poverty Measure: Its Core Concepts, Development, and Use - CRS Reports.* https://crsreports.congress.gov/product/pdf/R/R45031.