Datasheet for 'Systolic Blood Pressure (SBP) Dataset'*

INF312 Worlds Become Data: Tutorial 10

Mingjia Chen

March 19, 2024

For the purpose of this tutorial, the hypothetical datasheet is constructed based on the textbook by Wickham et al. (2019) and questions are extracted from Gebru et al. (2021), using open source statistically programming language R (R Core Team 2023). The dataset includes systolic blood pressure and other health factor, and is provided by University of Toronto Scarborough course called STAC67 Regression Analysis. I got to obtain and analyze the dataset for my last degree [1].

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 for speculating if individual health factors could contribute to influence one's systolic blood pressure. The size of the sample is relatively large (500 people), however, we do not know what the population the sample was drawn from.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was obtained from University of Toronto Scarborough Statistic department for studying purposes.
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.*
    - N/A

*Code and data are available at: https://github.com/MjChen120/INF312Tutorial10.git. Thanks to Catherine Punnoose for feedbacks.

[1]https://github.com/MjChen120/Past_Research/tree/main/R/STAC67_Research_Analysis_Paper

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.*

    - The dataset represents demographic backgrounds (gender, age, race, income, education), marital status, health related behaviors (exercise level, alcohol use, stress level, salt (NaCl) intake level, treatment (for hypertension)), and health characteristics (weight, height, Body Mass Index (BMI), overweight, childbearing potential). Most of the health responses are measured by medical measurement tools.

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

    - There are three types of instances in total: Four binarys, five continuous, and nine categorical.

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 contains a sample of 500 participants. Although having a large sample size, the data collection is not described as random selection and we do not know what population the sample was drawn from; hence, the sample is not representative of the large set.

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.*

    - Most numeric variable are raw, as they stayed unprocessed or cleaned yet for analysis. Body Mass Index (BMI) is calculated based on weight and height. The categorical and binary variables such as gender and cigarette smoking status are processed into numbers (1/2/3) or letters (Y/N,M/F) for further analysis.

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

    - Systolic Blood Pressure (SBP) is a continuous response variable recorded. Gender is a binary variable indicating whether the participant is male or female. Binary variables marital status, smoking status, and treatment for hypertension is in recorded as either yes (Y) or no (N). Age, weight, height, BMI are continuous variables

recorded. Overweight instance fits in three categories such as normal (1), overweight (2), and obese (3). Race and childbearing potential each have more than two categories. At last, exercise level, alcohol use, stress level, salt intake level, income level, and education level consists of three categories low (1), medium (2), and high (3).

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.*

    - The exact races assigned for each number is missing from this individual instance. For instance, we only know there are four number categories of race, the exact assignment is unavailable in the dataset description. It is probably an approach to minimize racial bias from researchers when dealing with the dataset.

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.*

    - The relationship between weight, height, and body mass index (BMI) was made exolicit with the calculation equation ((weight/height^2)*703) in the description.

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.*

    - N/A

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

    - N/A

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.*

    - The dataset is only available for studying purpose at the University of Toronto.

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.*

- N/A

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.*

    - N/A

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.*

    - Since ages and genders are in relatively equa distribution through out the dataset, there is no sub-populations identified.

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.*

    - N/A

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.*

    - Some of the health data is contained that might be considered sensitive such as marital status, smoking status, weight, height, BMI, overweight, race, alcohol use, stress level, childbearing potential and treatment for hypertension. In addition, financial data such as income level is also included in the dataset.

16. *Any other comments?*

    - N/A

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 for most of the instances were acquired through surveys. SBP responses were acquired through medical measurement tools.

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 manual recording.

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

   - N/A

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)?*

   - N/A

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.*

   - N/A

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.*

   - N/A

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

   - TBD

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.*

   - The individuals should be aware of the collection, however, this part of information is not included.

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.*

   - The individuals should consented on the used of their data, however, this part of information is not included.

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.*

    - N/A

12. *Any other comments?*

    - N/A

Recommended Uses

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

    - The dataset has been used to speculate if the individual health factors could influence one's Systolic Blood Pressure (SBP).

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 access point.*

    - https://github.com/MjChen120/Past_Research/tree/main/R/STAC67_Research_Analysis_Paper

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

    - N/A

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?*

    - The dataset is processed and recorded without the possibility of racial biases. The dataset does not have potential risks or harms.

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

    - Do not use it outside of the context of study purposes for University of Toronto due to intellectual property rules.

6. *Any other comments?*

    - N/A

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.