Datasheet for 'Exploring the Determinants of Body Weight: A Bayesian Analysis of Anthropometric and Demographic Factors Among Female U.S. Army Personnel'*

Huayan Yu

2024-12-11

Extract of the questions from Gebru et al. (2021).

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 investigate the relationship between body weight and anthropometric and demographic predictors among female U.S. Army personnel. It addresses the gap in research on female military personnel, particularly concerning their health, operational readiness, and body composition.
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 created by the Natick Soldier Research, Development, and Engineering Center (NSRDEC) on behalf of the U.S. Army as part of the 2012 ANSUR II survey.
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.*
    - The dataset was funded by the U.S. Department of Defense as part of its efforts to improve operational readiness and health monitoring among military personnel.
4. *Any other comments?*

*Code and data are available at: https://github.com/huayan1998/Exploring_the_Determinants_of_Body_Weight

- This dataset is one of the most comprehensive anthropometric surveys conducted on military personnel, focusing specifically on female soldiers.

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 instance represents an individual female U.S. Army soldier, with data points including their anthropometric measurements (e.g., height, weight, waist circumference) and demographic information (e.g., age, ethnicity, military component).

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

    - The dataset contains a total of 1,986 instances, each representing a female U.S. Army personnel. These instances include detailed anthropometric and demographic measurements.

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 of female U.S. Army personnel from various military branches. While efforts were made to ensure representation, it is not fully representative of the entire female military population due to limitations in geographic and occupational coverage.

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 raw anthropometric measurements (e.g., height, weight, waist circumference) and demographic features (e.g., age, ethnicity, and military component).

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

    - Yes, the dataset includes a target variable, weightlbs, which represents the body weight of each individual in pounds. This variable serves as the dependent variable in analyses aiming to understand the relationship between body weight and various anthropometric and demographic predictors.

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 missing information is reported, as missing values were removed during preprocessing.

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 explicit relationships between instances exist, as each instance represents an independent individual.

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.*

   - Yes, this research uses an 80/20 split, with 80% of the data used as the training set and 20% as the testing set. This approach ensures that the model is trained effectively on a majority of the data while retaining a portion for unbiased performance evaluation, as demonstrated by the comparison of RMSE between the full and reduced models.

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

   - Yes, the dataset may have sources of noise such as measurement variability due to inconsistencies in using instruments like calipers and tape measures, as well as missing values that can affect data completeness. These issues were addressed by standardizing measurements and removing rows with missing values during preprocessing to ensure accuracy and reliability.

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 largely self-contained and does not rely on external resources for its core variables, as all data were collected directly through the 2012 ANSUR II Female Dataset. While the dataset includes demographic information cross-referenced with administrative records, there are no dependencies on external, dynamically changing resources such as websites or social media. The dataset is publicly available and distributed under terms that do not impose restrictions, ensuring accessibility for

researchers without licensing fees or additional conditions. The archival version of the dataset, as it existed at the time of creation, is well-documented and maintained for consistency.

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, the dataset does not contain data that would typically be considered confidential. It includes anthropometric and demographic measurements collected from U.S. Army personnel under standardized protocols, but no sensitive personal identifiers or non-public communications are present. The data were anonymized to protect individual privacy, ensuring compliance with ethical and legal standards.

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, the dataset does not contain any offensive or distressing content.

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.*

    - Sub-populations are identified by demographic variables, such as age, ethnicity, and military component. These distributions are reported in the dataset documentation.

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.*

    - No, individuals cannot be identified directly or indirectly.

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.*

    - The dataset includes sensitive demographic data such as ethnicity and age, but these are anonymized and used solely for analysis.

16. *Any other comments?*

    - Incorporating additional lifestyle variables such as physical activity levels or dietary habits could enhance its utility and explanatory power in future research.

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.*

    - Anthropometric data (e.g., height, weight, waist circumference) was directly observed and measured using calibrated tools by trained personnel. Demographic data (e.g., age, ethnicity, military component) was self-reported and cross-verified against administrative records for accuracy.

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?*

    - Data was collected using calibrated instruments such as stadiometers, measuring tapes, and calipers for anthropometric measurements. Manual entry was used for demographic data, with verification against official military records to ensure reliability.

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

    - The dataset was derived using a stratified random sampling strategy to ensure representation across military components (Active Duty, Reserves, National Guard), age groups, and ethnicities.

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)?*

    - Data collection was conducted by trained personnel under the supervision of the Natick Soldier Research, Development, and Engineering Center (NSRDEC). Compensation details were not explicitly mentioned but likely followed standard military protocols.

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 collected as part of the 2012 ANSUR II survey conducted by the Natick Soldier Research, Development, and Engineering Center. The timeframe of data collection aligns with the creation of the dataset, as all measurements and associated demographic information were recorded during the survey period in 2012 using standardized protocols. There is no mismatch between the data collection and creation timeframes.

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.*

    - The dataset originates from the 2012 ANSUR II survey, conducted by the Natick Soldier Research, Development, and Engineering Center, which followed rigorous ethical standards for data collection. Although specific details about an institutional review board (IRB) review are not explicitly mentioned, the data collection adhered to military protocols ensuring participant confidentiality and informed consent. These procedures are consistent with ethical guidelines for research involving human subjects in military contexts. Supporting documentation, such as the official survey reports, may provide further details and can be accessed through sources like the Natick Soldier Research Center or related publications.

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

    - The data was obtained by me from the platform Data World, where the 2012 ANSUR II survey dataset is publicly available.

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.*

    - Participants were likely informed as part of the data collection process conducted under military protocols, though specific notification details are not provided.

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.*

    - Consent was likely obtained in accordance with standard military data collection procedures, though specific consent forms or language are not available.

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).*

    - No explicit mechanism for revoking consent is mentioned, as the data collection was part of military operations.

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 specific data protection impact analysis is mentioned, but the data is anonymized to protect individual identities.

12. *Any other comments?*
    - No.

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.*
    - Yes, preprocessing was performed, including removing missing values, renaming variables for consistency, and scaling measurements (e.g., converting waist circumference from millimeters to centimeters).
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.*
    - The raw data was retained in its original form for future reference and potential reanalysis.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - The preprocessing was performed using R, and the scripts are available in the project repository.
4. *Any other comments?*
    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - No, the dataset has not been used for any specific tasks prior to this analysis. It was obtained for the purpose of exploring the relationship between body weight and key anthropometric and demographic predictors.
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.* -No, there is no repository that links to papers or systems that use this dataset. It was obtained from Data World and does not have an associated archive for tracking its usage in research or systems.
3. *What (other) tasks could the dataset be used for?*

- The dataset could be used for health monitoring, ergonomic research, fitness program optimization, and studies on gender-specific anthropometric trends.

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?*

    - Consumers should be cautious about generalizing findings beyond female U.S. Army personnel, as the dataset is not representative of broader populations.

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

    - The dataset should not be used for identifying individuals or making decisions that could unfairly impact subgroups without additional validation.

6. *Any other comments?*

    - No.

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.*

    - Yes, the dataset is publicly available and can be accessed by researchers or institutions interested in anthropometric studies or military health monitoring.

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

    - The dataset is distributed via Data World (https://data.world/datamil/ansur-ii-female). It does not have a DOI.

3. *When will the dataset be distributed?*

    - The dataset has already been distributed and is publicly accessible.

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.*

    - The dataset is distributed under a public license with terms outlined on the Data World platform. There are no fees associated with its use for research purposes.

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, there are no IP-based or other restrictions imposed by third parties on the data associated with the instances. The dataset is publicly available on the Data World platform without licensing terms or fees restricting its use.

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 export controls or regulatory restrictions apply to the dataset.

7. *Any other comments?*

    - The dataset is easily accessible and designed for educational and research purposes.

Maintenance

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

    - The dataset is hosted and maintained by Data World and the original creators at the Natick Soldier Research, Development, and Engineering Center.

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

    - The dataset curators can be contacted through the Data World platform.

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

    - No formal erratum has been issued.

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)?*

    - The dataset is static and unlikely to receive updates. Any changes would be communicated through the Data World platform.

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.*

    - There are no explicit retention limits, as the dataset is anonymized and publicly available.

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.*

    - The dataset is not expected to undergo major changes, so the current version will remain accessible indefinitely.

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.*

    - There is no formal mechanism for external contributions to this dataset.

8. *Any other comments?*

    - The dataset is intended as a standalone resource for academic and research purposes, with no current plans for collaborative updates.

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.