Depression and Associated Risk Factors: NHANES Data Analysis

02051768

Compiled: May 12, 2025

Github Repo: https://github.com/yichen-hazel/MATH70076-CW2/releases/tag/v1.0.0

1 Project Description (approx. 250 words)

This project explores the relationship between depression levels and a range of influencing factors using publicly available NHANES data. We focus on survey years 2015–2018 and 2021–2023 with seven modules including DEMO, DPQ, ALQ, SMQ, PAD, DIQ, and MCQ. The outcome of interest is the sum of PHQ-9core from the DPQ module, which is widely used to test the depressive symptoms. Our aim is to investgates how demographic, behavioral, and health-related variables are associated with mental health outcomes.

We carefully match variable formats across different survey years. Missing values are either classified as a separate group, removed from the analysis, or filled in using appropriate methods depending on the context. After merging and cleaning the datasets, we conduct exploratory data analysis (EDA) on both numerical and categorical variables through summary tables and a various type of plots. We examine distributions, missing patterns, and potential correlation trend. Then the generalized linear model are used for discrete data in the this report. Several model types are considered here, including Poisson, quasi-Poisson, and negative binomial regression, to deal with the potential overdispersion.

Variables are selected base on statistical significance and model comparison using AIC. We further explore interaction terms to capture modifying effects. Model diagnostics including residual analysis and goodness-of-fit checks are conducted to evaluate performance. Coefficients are interpreted in terms of incidence rate ratios, with attention to practical significance and confidence intervals.

We also acknowledge several limitations of our analysis, such as missing data and the cross-sectional nature of the NHANES dataset. Possible directions for future work include using classification or binary outcome models, exploring longitudinal data if available, and incorporating survey weights to improve representativeness.

2 Assessment Criteria

In 2-3 bullet points each and at most 1 page in total, describe how your submission addresses each of the assessment criteria below. You may delete this italicised text when filling in the template.

Technical Competence: Proficiency in data collection, processing, analysis, and coding.

- Combined and cleaned data from multiple modules and years.Matched variable formats and handled missing values.
- Performed EDA and fitted count models to address overdispersion in PHQ-9 scores.

User Interface: Design, functionality, and usability of the final data product.

- Report is clearly structured and easy to follow.
- Helps readers understand depression risk factors and identify vulnerable groups

Analysis and Interpretation: Depth of analysis, appropriate use of statistical methods, and meaningful interpretation.

- Compared several generalized linear models and selected the best based on AIC. Added interaction terms informed by EDA.
- Interpreted all model coefficients in terms of direction, size, and significance

Presentation and Communication: Clarity, organisation and effectiveness of written and visual communication.

- Used ggplot2 with contrasting colors to highlight key effects. Formatted tables using knitr and kableExtra for better readability
- Adjusted headings, fonts, and figure styles for clarity.

Reproducibility and Documentation: Clarity and completeness of documentation for product use and reproducibility.

- Saved raw and processed data separately. Included NHANES variable definitions and a brief guide for the combined dataset
- Wrote separate R scripts for each part of the workflow.
- Saved all plots used in analysis and reporting.

Project Management: Considered and effective use of project management and version control systems.

- Scripts are modular and organized by task.
- Final results and report are published on GitHub with a tagged release.

3 Project Reflection

Reflect on the experience of creating your data product. In 6 bullet points and at most 1 page total, summarise the following.

- *3 things you have learned as part of this process,*
- *2 aspects of the project that you found challenging or would approach differently with hindsight,*
- *1 aspect of the project that you would like to learn more about in the future.*

You may delete this italicised text when filling in the template.

Learnings:

- How to combine and clean large-scale raw survey and questionnaire data across multiple NHANES modules and cycles, and how to select appropriate variables for analysis.
- Get more familiar with the full process of developing a model for overdispersed count data, including model selection, evaluation, and interpretation.
- How to build a clearly structured and highly reproducible project from data preparation to reporting.

Challenges:

- It was challenging to decide whether a variable that seemed highly relevant was worth keeping when its data was incomplete or unclear.
- Understanding and interpreting interaction terms was more complex than expected, especially when the effects varied across subgroups and the interaction itself was hard to explain.

Further Development:

- We can explore longitudinal survey data in the future to better understand how depression risk changes over time.