Final Project Report

STATEMENT OF RESEARCH QUESTION

Dependent/Outcome Variable: depression_score (score of depression severity based on the PHQ-9 questionnaire)
Independent/Predictor Variables (Sociodemographic Factors): age, race/ethnicity, marital status, education level, income to poverty ratio, health insurance status, if they saw a mental health professional in the last year (mental_health_seen), employment status, and food security status.
Goal of my model is to find the relationship and association between sociodemographic factors associated with depression severity among women of reproductive age in the NHANES dataset.

MOTIVATION AND BACKGROUND

I developed an interest in this topic after studying mental health and its risk factors during undergrad. Though I couldn't explore it deeply at the time, it has remained an area of interest. Depression is a significant public health concern, particularly among women of reproductive age, as it can negatively impact both maternal and child health[1]. Sociodemographic factors such as age, education, marital status, income, and employment are key predictors of mental health disparities[2]. Using the nationally representative NHANES dataset, I aim to conduct a robust quantitative assessment of the associations between these factors and depression severity among women of reproductive age with improved generalizability due to the dataset's large sample size.

1 EXPLORATORY ANALYSIS AND QUALITY CONTROL OF THE DATA

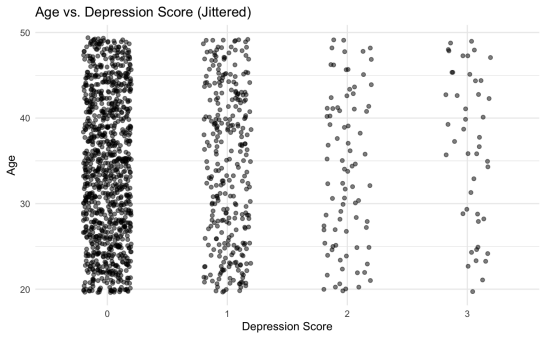
1.1 Data Transformation + Assessing/removing/imputing missing values

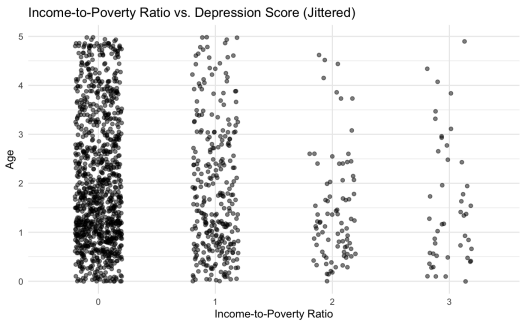
Variable	**Meaning**	**Justification for Filtering**
depression_status	Depression severity: 0 = not difficult 1 = somewhat difficult 2 = very difficult 3 = extremely difficult	This describes the "difficulty caused" by depression, relating to severity. Removed all "refused", "NA", "IDK" values
age	Ages from 0 to 79.	Reproductive age women are 15-49[1]. Due to disclosure risks, marital status is only released for persons 20+ age and decided to keep the range as 20-49 instead
gender	1 = male 2 = female	only kept 2 = female because I am looking at reproductive age women
race_ethnicity	1 = Mexican American 2 = Other Hispanic 3 = Non-Hispanic White 4 = Non-Hispanic Black 6 = Non-Hispanic Asian 7 = Other Race - Including Multi-Racial)	Removed all missing values

marital_status	1 = Married/Living with Partner 2 = Widowed/Divorced/Separated 3 = Never married	Removed all "refused", “NA”, “IDK" values.
education_level	1 = Less than 9th grade 2 = 9-11th grade (includes 12th grade with no diploma) 3 = high school graduate/GED or equivalent 4 = some college or AA degree 5 = college graduate or above	Removed all "refused", “NA”, “IDK" values
income2poverty_ratio	Ranges from 0 to 4.98	Omitted values at or above 5.00 (they were coded as 5.00 or more because of disclosure concerns
health_insurance	1 = yes 2 = no	Removed all "refused", “NA”, “IDK" values
mental_health_seen	1 = yes 2 = no	Removed all "refused", “NA”, “IDK" values
employment_status	1=Working at a job or business 2 = With a job or business but not at work 3 = Looking for work 4 = Not working at a job or business	Removed all "refused", “NA”, “IDK" values. Question is asking respondents if they were at work in the past week. So, I combined 1 and 2 to reflect that the individual is employed.
food_security_status	1=AD full food security 2=Household marginal food security, 3=AD low food security 4=AD very low food security	Removed all "refused", “NA”, “IDK" values

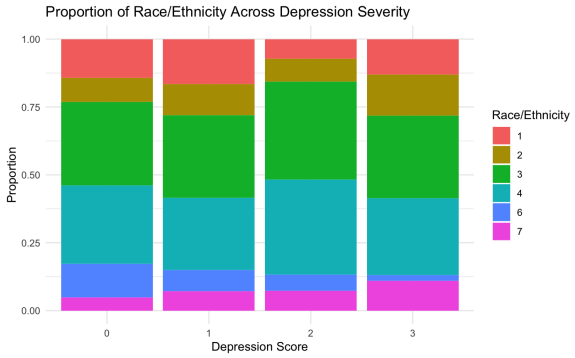
1.2 Plots showing relationship between variables of interest.

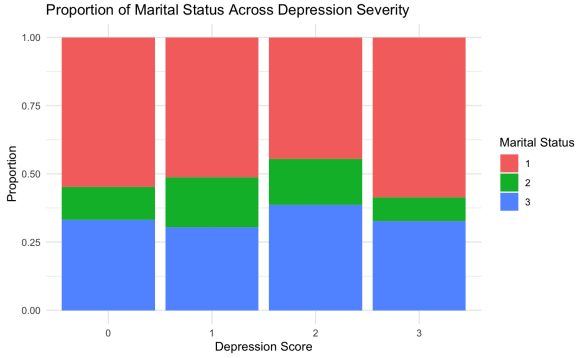
My variables were primarily of two types: numerical and categorical (ranked). For numeric data (age and income-to-poverty ratio), I used jittered scatter plots to visualize the distribution across each depression score.

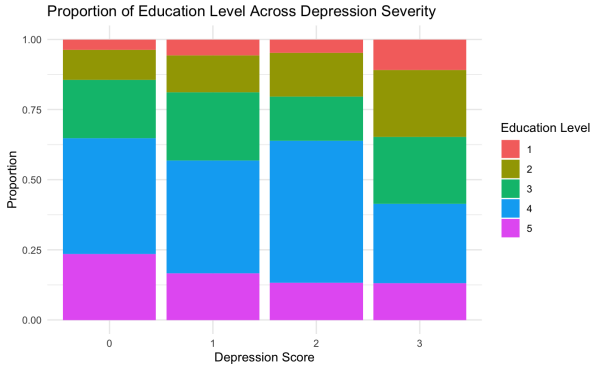


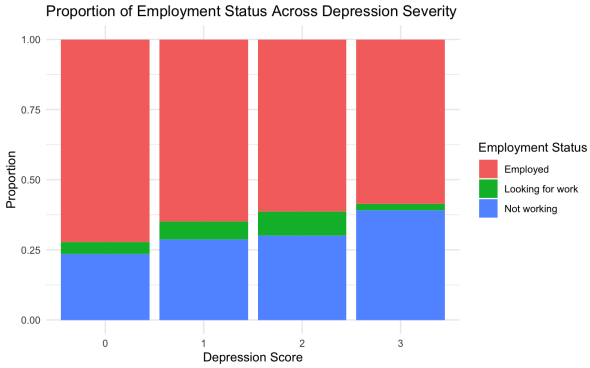


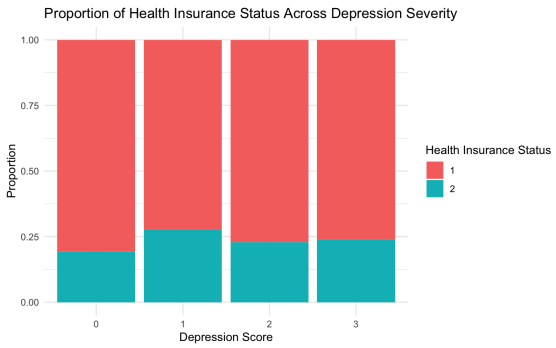
For the categorical variables, I used stacked bar plots to visualize the proportions across each depression score.

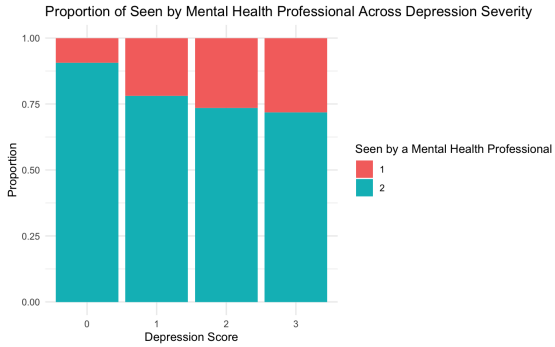


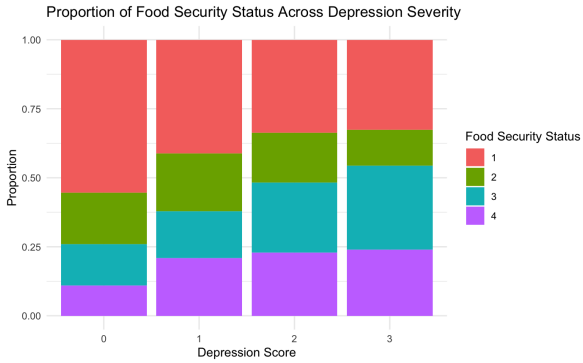












1.3 Check correlations/collinearity:
I conducted a basic Spearman correlation and found that most variables had no significant correlation, however, there were slight correlations with income2poverty_ratio, employment_status, mental_health_ seen, and food_security_status. I considered a rho value close to 0 as indicating no significant correlation, with values closer to -1 or 1 representing a stronger correlation. My interpretation for each variable is written as a comment in my R code.

1.4 Scale predictors as needed:
I only scaled my numeric variables (age, income2poverty_ratio)

1.5 Code categorical variables appropriately:
Although I used factor() when constructing my plots, I also ensured that my variables were properly converted to factors

2: CHOOSE A REGRESSION OR CLASSIFICATION MODEL
2.1 Model Chosen:
ordinal logistic regression model

2.2 Evaluating Assumptions of Model:
The biggest "issue" with my data was that I had numeric values, binary predictors (yes,no), and multi-level categorical variables (e.g. 1,2,3). The numeric predictors can be included directly in the model, the binary treated as categorical variables coded as 0/1, and the categorical variables being treated as ordinal predictors.

I am assuming that the ordinal categories of the depression_score and the independent variables have a meaningful structure and order, and that each independent variable is measured at an appropriate level. I assume that the observations in my dataset are independent of one another. I also assume there is no multicollinearity, as the correlation analysis in Section 1.3 indicated that any correlations present were not strong. Based on the exploratory plots, no extreme outliers were observed, so I assume that outliers are not a concern in this dataset.

3: RUN THE MODEL

3.1 Run Initial Model:
I ran the ordinal regression model including all independent variables.

3.2 Variable Selection:
I performed stepwise variable selection (AIC) to identify the best predictors for model fit while balancing complexity. Initially, all variables were included. Through the stepwise process, variables were added or removed based on their impact on the AIC. Marital status was the first variable to be removed as it decreased the AIC the most. The final model included income2poverty_ratio, health_insurance, mental_health_seen, and food_security_status as predictors.

3.3 Re-Run Model (Based on Variable Selection):
Using the selected predictors, I updated the model and ran it with the four variables identified through the stepwise process.

4: EVALUATE MODEL FIT
4.1 Choose, calculate, and appropriately interpret a metric
Based on the output from my model in 3.3, I will interpret the coefficients. The coefficients represent the change in the predicted likelihood of the outcome for each one-unit increase in the predictor variable.

4.2 Justify the metric you chose based on the goal of your model
The variables and their coefficients are: income2poverty_ratio (-0.1118), health_insuranceNo (0.3585), mental_health_seenNo (-1.0006), food_security_statusMarginal, Food Security (0.2610), food_security_statusLow, Food Security (0.5457), and food_security_statusVery Low Food Security (0.7475).

My interpretation for each coefficient is as follows:

- As the income-to-poverty ratio increases, the likelihood of reporting higher depression severity decreases.
- Individuals without health insurance are more likely to report higher depression severity.
- People who have not seen a mental health professional in the last year are more likely to report higher depression severity.
- The more severe the food insecurity, the stronger the association with higher depression severity.

5: COMPARE YOUR MODEL TO ONE WITH DIFFERENT VARIABLES

5.1 Justify alternate model chosen
I decided to remove the variables health_insurance and mental_health_seen from the model because these factors seem like more “obvious” reasons (i.e health access barriers) for why any given person would have more severity in depression. With this, I aimed to explore the social and economic factors.

5.2 Run alternate model

5.3 Calculate a metric to compare models
To compare the models, I used a likelihood ratio test to assess the goodness-of-fit between a more complex model and a simpler one. This metric helps determine if the more complex model fits the data significantly better. I found it useful when comparing the models in 3.3 and 5.2. The likelihood ratio statistic was -27.32453, which indicated a difference in model fit. However, with a p-value of 1 (greater than 0.05), this difference is not statistically significant. Therefore, the model in 5.2 does not fit better than the model in 3.3, and adding or removing the specified predictors does not improve the model.

6: CONDUCT ONE FOLLOW UP ANALYSIS

6.1 Choose an analysis which fits with the over goal of your model

I decided to do a regression at a different age group and see the association of depression severity and income2poverty_ratio, food_security_status, marital_status, and education_level among individuals aged 34 or younger. I chose 34 because you are considered to have a "geriatric pregnancy" if you are 35 or older[3].

6.2 Appropriately interpret the results

Looking at the coefficients that are statistically significant (based on t-value). food_security_statusLow Food Security = 0.5740. This indicates that Individuals with low food security have a significantly higher likelihood of reporting higher depression severity compared to those with full food security. food_security_statusVery Low Food Security = 0.7586). This indicates that individuals with very low food security are more likely to report higher depression severity compared to those with full food security.marital_statusWidowed/Divorced/Separated = 0.6384.This indicates that being widowed, divorced, or separated is associated with a higher likelihood of reporting higher depression severity compared to being married or living with a partner.

Overall, food security status is the most consistent and statistically significant predictor of depression severity. With the association showing that as food security status worsens, depression severity increases. To add, marital status provided some unique insights as the Individuals who are widowed, divorced, or separated are more likely to experience higher depression severity compared to those who are married or living with a partner. Also, this model indicated no significant association between education level, income-to-poverty ratio and depression severity.

References

1. Guo, N., Robakis, T., Miller, C., & Butwick, A. (2018). Prevalence of Depression Among Women of Reproductive Age in the United States. *Obstetrics and gynecology,* 131(4), 671–679. https://doi.org/10.1097/AOG.0000000000002535
2. Carpena, M. X., Dumith, S. C., Loret de Mola, C., & Neiva-Silva, L. (2019). Sociodemographic, behavioral, and health-related risk factors for depression among men and women in a southern Brazilian city. Revista brasileira de psiquiatria (Sao Paulo, Brazil : 1999), 41(5), 396–402. https://doi.org/10.1590/1516-4446-2018-0135
3. Pregnancy at Age 35 Years or Older: ACOG Obstetric Care Consensus No. 11. (2022). *Obstetrics and gynecology,* 140(2), 348–366. https://doi.org/10.1097/AOG.0000000000004873

Data Used:

1. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_DEMO.htm#Data_Processing_and_Editing
2. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_DPQ.htm#Data_Processing_and_Editing
   https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_OCQ.htm#Data_Processing_and_Editing
3. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_HIQ.htm#Data_Processing_and_Editing
4. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_HUQ.htm#Data_Processing_and_Editing
5. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_FSQ.htm#FSDAD