National Health Survey Data Analysis: Lifestyle Impact on Health Outcomes

A Logistic Regression Study on Smoking, Diet, and Mental Health

STUDENT NAMES

- Ryan Yahnker
- Kian Jadbabaei
- Spencer Kung

1 Abstract

This study investigates three health-related hypotheses using data from a national health survey. Specifically, we examined whether

1. The presence of a liver condition influences the quality of a person's diet.
2. Increased cigarette consumption (measured by smoke_amount) is associated with higher odds of reporting a COPD diagnosis.
3. Insufficient sleep on weekdays (6 hours) correlates with greater likelihood of experiencing frequent depressive symptoms.

To test these hypotheses, we employed ordinal logistic regression for the first and third hypotheses, given their ordinal outcome variables. For the second hypothesis, we used binary logistic regression to model the relationship between continuous smoking behavior and the binary presence of COPD. Our findings revealed a statistically significant association between higher smoke_amount and increased odds of reporting COPD, supporting Hypothesis 2. However, no statistically significant relationships were found for Hypotheses 1 and 3. These results underscore the health risks associated with smoking while highlighting the need for further investigation into the effects of liver conditions and sleep on diet and mental health outcomes, respectively.

2 Introduction

The dataset we are working with comes from the National Health and Nutrition Examination Survey, a survey that collects detailed health information from a representative sample of adults in the United States. It includes a wide range of variables that cover background information, medical conditions, lifestyle habits, and self-reported health outcomes. For this project, we selected variables that allow us to explore connections between physical and mental health, with a focus on three specific research questions.

The **first question** asks whether having a liver condition influences how individuals rate the quality of their diet.

The **second question** looks at whether smoking intensity, measured by the average number of cigarettes smoked per day, is associated with the likelihood of reporting a COPD diagnosis.

The **third question** examines whether the number of hours someone sleeps on weekdays relates to how often they report experiencing depressive symptoms.

Below are the formal hypotheses tests, corresponding to our three research questions.

Hypothesis 1: Liver Condition and Diet Quality

We examine whether having a liver condition affects how individuals rate their diet quality (ordinal outcome).

H_0: The distribution of diet quality ratings is the same for individuals with and without a liver condition.
H_A: The distribution of diet quality ratings differs between individuals with and without a liver condition.

Hypothesis 2: Smoking and COPD Diagnosis

We test whether higher cigarette consumption is associated with increased odds of reporting a COPD diagnosis (binary outcome).

H_0: There is no association between smoking amount and the likelihood of a COPD diagnosis.
H_A: Higher smoking amount is associated with increased likelihood of a COPD diagnosis.

Hypothesis 3: Sleep Duration and Depression Severity

We investigate whether shorter weekday sleep is related to more frequent depressive symptoms (ordinal outcome).

H_0: There is no difference in depression frequency between individuals with short sleep and those with adequate sleep.
H_A: Individuals with short sleep report more frequent depressive symptoms compared to those with adequate sleep.

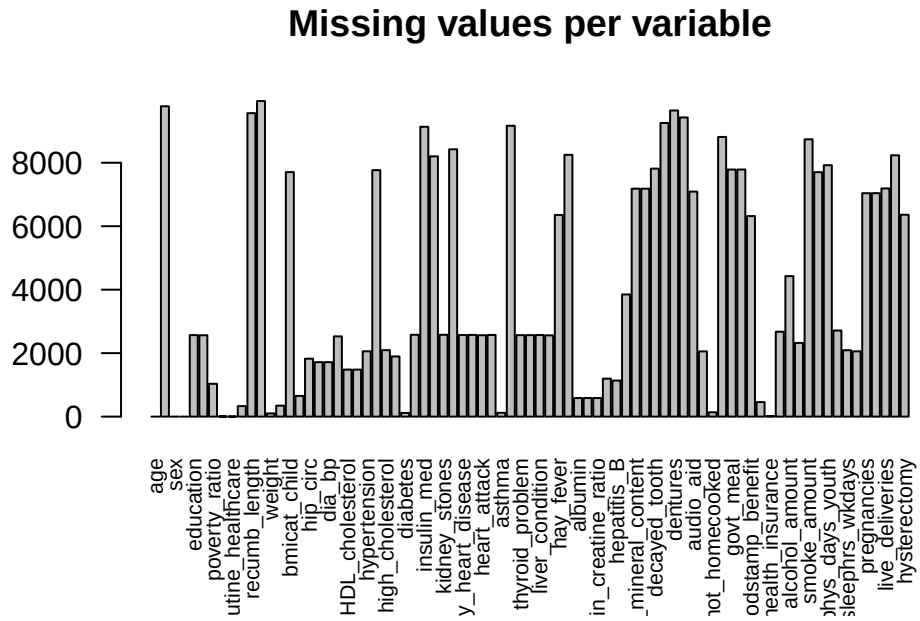
These questions and their respective hypotheses are grounded in the idea that behaviors and health conditions are often related in meaningful ways. Each research question is tied to a hypothesis that we investigate through exploratory data analysis. This involves preparing the data, calculating descriptive statistics, and creating visualizations to identify trends and potential associations between variables. Our goal is to develop a deeper understanding of how these health factors interact within the NHANES population and to generate insights that could guide future analysis.

3 Data Processing

3.1 Data Cleaning

To prepare the dataset for analysis, we cleaned and filtered the NHANES dataset based on the variables relevant to our three research questions. Below, we outline each step in our data cleaning process.

We first explored missing values by counting the number of NAs in each column and visualizing them in a barplot.



We found that no row was fully complete across all 79 columns, as most participants were missing at least one variable. Therefore, dropping all rows with any missing value would drastically reduce the dataset.

```
[1]  0 79
```

Rather than dropping rows across the full dataset, we limited our cleaning to the variables directly related to our three hypotheses (age, liver_condition, diet_survey, COPD, smoke_amount, sleephrs_wkdays, and depression).

Some variables, such as smoke_amount, were stored as character strings with non-numeric characters. We cleaned this variable to retain only numeric values

3.2 Identify outliers

To identify outliers in the `smoke_amount` variable, we used the **Interquartile Range (IQR) method**, a standard technique for detecting extreme values in a distribution.

3.2.1 Calculate the IQR

Using the IQR, we computed the lower and upper bounds:

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR} = -17.5$$

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} = 42.5$$

We then checked for any `smoke_amount` values falling outside these bounds.

Our lower smoke bound is -17.5 and our upper smoke bound 42.5.

```
[1] age          sex          smoke_amount
<0 rows> (or 0-length row.names)
```

After removing observations with missing values for our variables on interest, we find no outliers for our smoke_amount which is the only numeric variable of interest.

Finally, to prepare for modeling and visualization, we performed several final transformations:

- Created a binary sleep_category variable to distinguish between “Short sleep” (6 hours) and “Adequate sleep” (>6 hours).
- Converted ordered factors like depression and diet_survey into numeric values when appropriate, allowing correlation analysis and regression modeling.

These transformations ensured variables were in formats appropriate for analysis and improved interpretability across plots and models.

4 Modeling Process

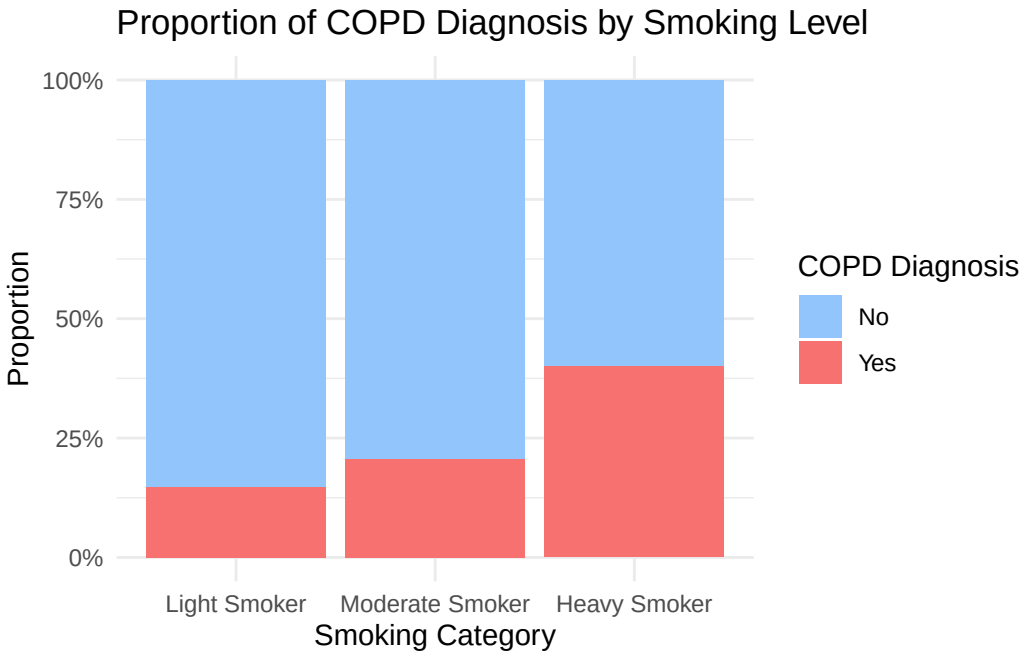
Our goal was to examine relationships between selected health behaviors and outcomes using models suited to the type and structure of each variable. For questions where the outcome was ordinal (like diet quality or depression frequency), we used ordinal logistic regression via the polr() function from the MASS package. This allowed us to preserve the ordered structure of the response categories. For the question where the outcome was binary (COPD diagnosis), we used binary logistic regression via the glm() function with family = binomial.

We chose these models because they align with our research questions and the nature of our data. Each model includes one main predictor based on our hypothesis, and we interpreted the coefficient and p-value to assess whether there is a statistically significant association.

5 Results

Table 1: Summary of Model Coefficients and p-values

Research.Question	Model.Type	Coefficient	p.value
Liver Condition → Diet Quality	Ordinal Logistic Regression	-0.0905	0.691
Smoking Amount → COPD Diagnosis	Binary Logistic Regression	0.0567	0.000
Weekday Sleep Hours → Depression Level	Ordinal Logistic Regression	0.0363	0.281



To better visualize the relationship between smoking and COPD, we grouped smoke_amount into categorical levels: Light, Moderate, and Heavy smokers. The plot shows a clear trend where higher smoking categories are associated with a greater proportion of COPD diagnoses, especially among heavy smokers.

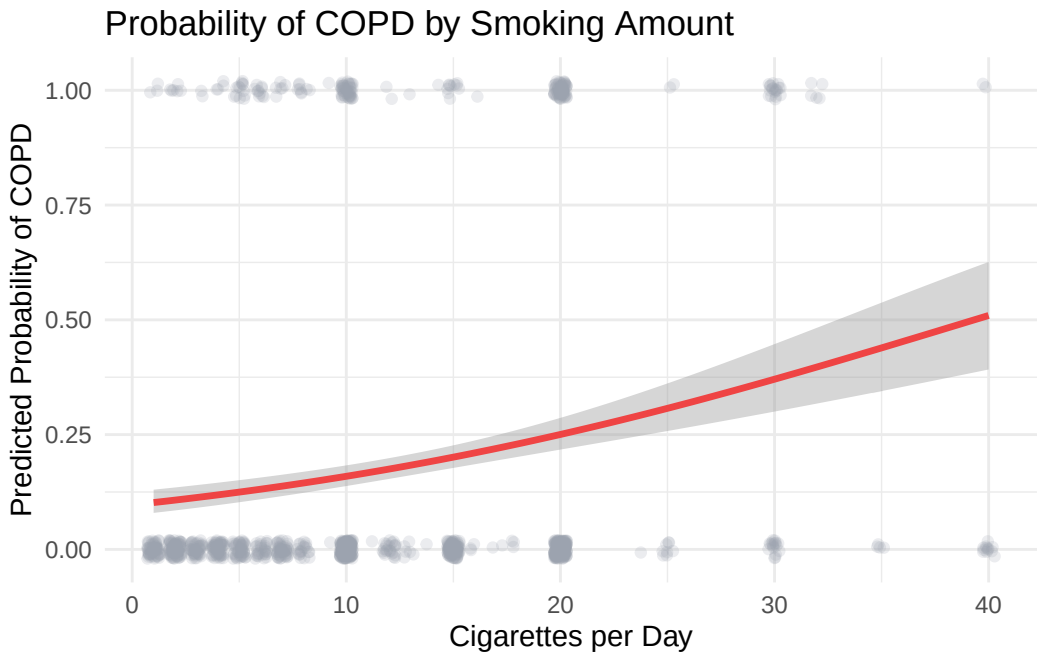
6 Interpretation

Our first research question asked whether individuals with a liver condition rate their diet differently than those without. The hypothesis was that the presence of a liver condition would influence diet quality, possibly encouraging healthier eating habits. However, the results from the ordinal logistic regression model showed no statistically significant relationship between liver condition and self-reported diet survey scores ($p = 0.691$). This suggests that, at least within this dataset, having a liver condition does not strongly impact how someone evaluates the quality of their diet. This might be because dietary change isn't consistently adopted after diagnosis, or because the self-rated diet score doesn't capture specific dietary adjustments tied to liver health.

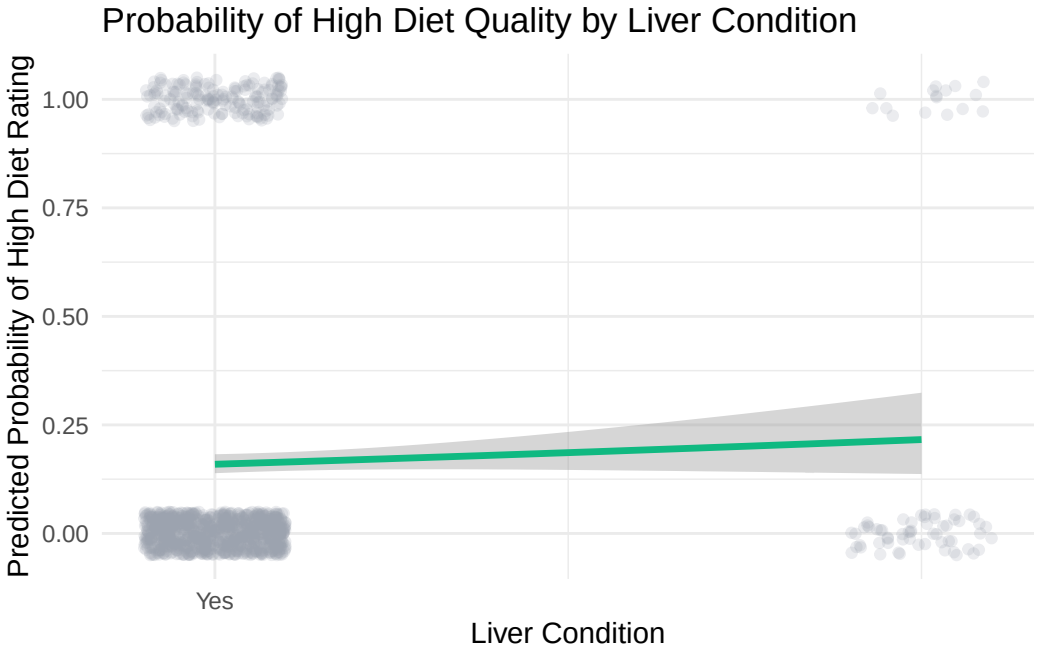
The second research question focused on whether the number of cigarettes smoked per day is associated with the likelihood of reporting a COPD diagnosis. The hypothesis was that higher smoking intensity would be linked to increased odds of having COPD, and this was clearly supported by the data. The logistic regression returned a statistically significant result ($p < 0.001$), and the coefficient was positive, meaning that as cigarette consumption increases, the odds of reporting COPD go up. This trend is also visible in the proportional bar plot, where the percentage of individuals with COPD rises steadily from light smokers to heavy smokers. These results align with well-established clinical evidence connecting smoking to chronic respiratory conditions, and they reinforce the strength of this relationship within the NHANES sample.

Our third research question explored whether weekday sleep duration is related to depressive symptom frequency. The hypothesis was that shorter sleep would be associated with more frequent symptoms of depression. Although the direction of the coefficient was positive, the ordinal logistic regression model did not return a significant result ($p = 0.281$). This means we do not have enough evidence to say that sleep duration predicts depression levels in this dataset. There are several possible reasons for this, including the fact that both sleep and depression were self-reported, which introduces variability, or that other factors like stress, physical health, or medication use may be more relevant in explaining depression symptoms than sleep alone.

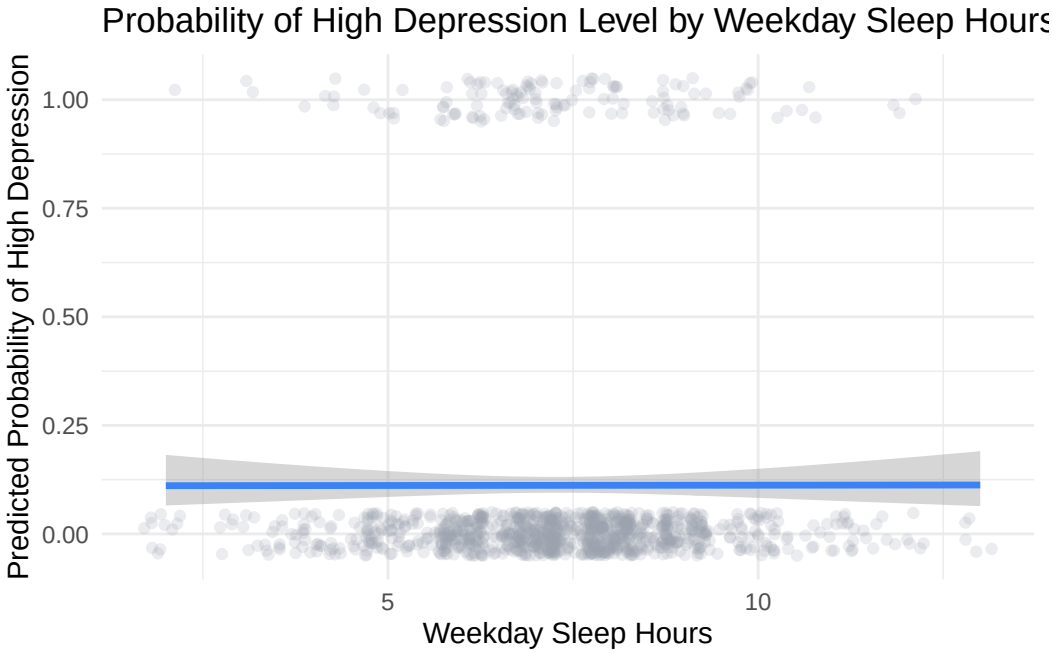
7 Visualization and Communication



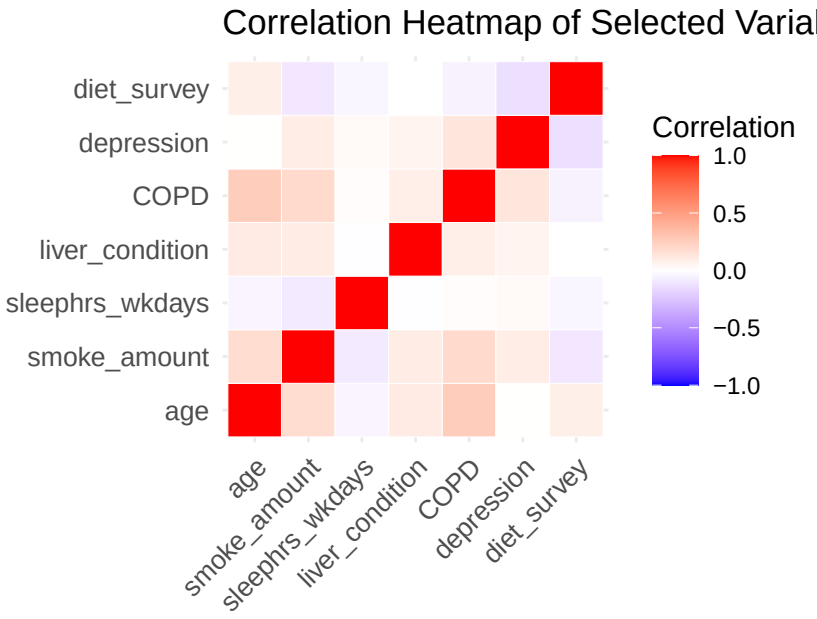
This plot shows the predicted probability of having COPD based on the number of cigarettes smoked per day. Each dot represents an individual, with 0 indicating no COPD and 1 indicating a COPD diagnosis. The red line shows the fitted logistic regression curve, and the shaded area represents the confidence interval. As smoking amount increases, the probability of having COPD also increases. The upward curve supports our earlier finding that smoking intensity is significantly associated with higher odds of COPD.



This plot shows the predicted probability of rating one's diet as "Very good" or "Excellent" based on liver condition status. The flat curve and wide confidence band indicate little to no difference in high diet quality ratings between those with and without a liver condition.



This plot shows the predicted probability of experiencing high levels of depression based on weekday sleep hours. The nearly flat curve and wide confidence band suggest no meaningful relationship between sleep duration and reported depressive symptoms.



This correlation heatmap shows generally weak linear relationships among the selected variables. Most of the off-diagonal cells are close to white, indicating low correlation. While smoking amount has a mild positive correlation with COPD and age, variables like diet quality, depression, and sleep hours show little to no strong association with the others. This supports the decision to model each outcome separately rather than relying on multicollinearity.

8 Conclusion and Recommendations

Out of the three relationships we tested, only one showed strong evidence of a meaningful association. Smoking amount was significantly linked to higher odds of COPD, and that relationship was consistent across every visualization and supported by the logistic model. The other two questions, whether liver condition impacts diet quality and whether sleep hours predict depression, did not show statistically significant results. That does not mean there is no relationship at all, but it means we could not detect one in this dataset.

If we were giving recommendations based on this analysis, the takeaway is clear. Smoking is a strong risk factor for COPD, and prevention efforts should continue to focus on reducing smoking behaviors. The lack of strong signals in the other two models suggests that improving diet or addressing depression may be more complex and influenced by additional factors not captured here.

There are a few limitations to consider. The data is self-reported, which introduces bias, especially in variables like diet and depression. We also simplified some outcomes into binary variables for visualization and modeling, which may lose important detail.

In future work, it would be helpful to include more context such as medical history, medication use, or stress levels to better understand what influences these outcomes. Using more advanced modeling approaches or interaction terms could also help uncover patterns that simple models might miss. It would also be beneficial to fit models including more predictors to uncover additional insights, rather than simply examining correlations between specific variables.