Logistic Regression Analysis for Predictors of Chronic Fatigue Syndrome

Phillip Braun

April 6, 2025

1 Introduction

Chronic Fatigue Syndrome (CFS), also referred to as myalgic encephalomyelitis (ME), is a disorder characterized by extreme fatigue that cannot be explained by any underlying medical condition. Often, it is accompanied by a variety of other symptoms such as sleep disturbances, cognitive dysfunction, and pain [3]. Identifying factors associated with CFS is critical for developing effective interventions and public health strategies. However, CFS is relatively rare and thus presents challenges for statistical modeling and inference.

Several studies have modeled the occurrence of CFS using both classical and modern statistical methods. Logistic regression has been used to identify predictors of CFS, including previous infection history and psychosocial factors [1, 2]. Machine learning approaches such as support vector machines and decision trees have also been used to classify CFS based on genomic data [4].

In this project we uncover multiple significant relationships between CFS and various demographic, lifestyle, and health-related predictors. This is done using a publicly available dataset from Kaggle titled "Healthcare Survey Data" [5] which contains preprocessed data from the Canadian Community Health Survey (2019). A logistic regression model is used, with the presence of CFS as the binary response variable. This model is not only used to determine significant predictors, but can also be used for classification purposes. Given the complexity of health data and the rarity of CFS, we focus on handling class imbalance and selecting an interpretable model. We applied a range of techniques, including stepwise selection via AIC, LASSO regularization, and weighted logistic regression.

The final weighted model revealed that older age, female gender, the presence of anxiety disorders and sleep apnea, higher stress levels, and lower mental health ratings were significantly associated with greater odds of reporting CFS. Lifestyle factors such as lower fruit and vegetable consumption and recent cannabis use also exhibited an association with CFS, although less significant.

Model evaluation using ROC analysis gave an AUC of 0.842, which is evidence of strong discriminative performance. However, the default thresholds led to poor specificity. Thus, we performed threshold tuning. To improve classification of the minority class, we selected the optimal

classification threshold using Youden's J statistic [7], defined as:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

This metric balances sensitivity and specificity, and identifies the threshold that maximizes true positive and true negative rates. This improved balanced accuracy to 76.8%, offering a more practical trade-off between sensitivity and specificity. Residual diagnostics confirmed that the model has good fit, with a few potentially influential observations.

2 Summary Statistics and Data Visualization

We begin by exploring the distribution and the structure of the data. Table 1 presents statistics for the predictors, the majority of which are categorical.

Variable	**Mean / %**	**SD / Count**
Age (18–34 / 35–49 / 50–64)	34.8% / 31.2% / 34.0%	–
Gender (Female)	51.3%	–
Anxiety Disorder (Yes)	15.2%	–
Sleep Apnea (Yes)	10.4%	–
Fruit and Vegetable Intake	173.4	60.2
Cannabis Use (Yes)	20.1%	–
CFS Prevalence (Yes)	1.1%	–

Table 1: Summary statistics for selected predictors and outcome

To visualize relationships between predictors and CFS, multiple figures were generated.

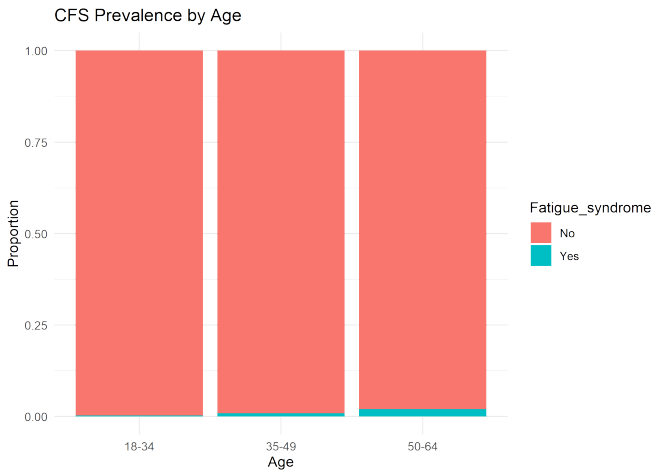
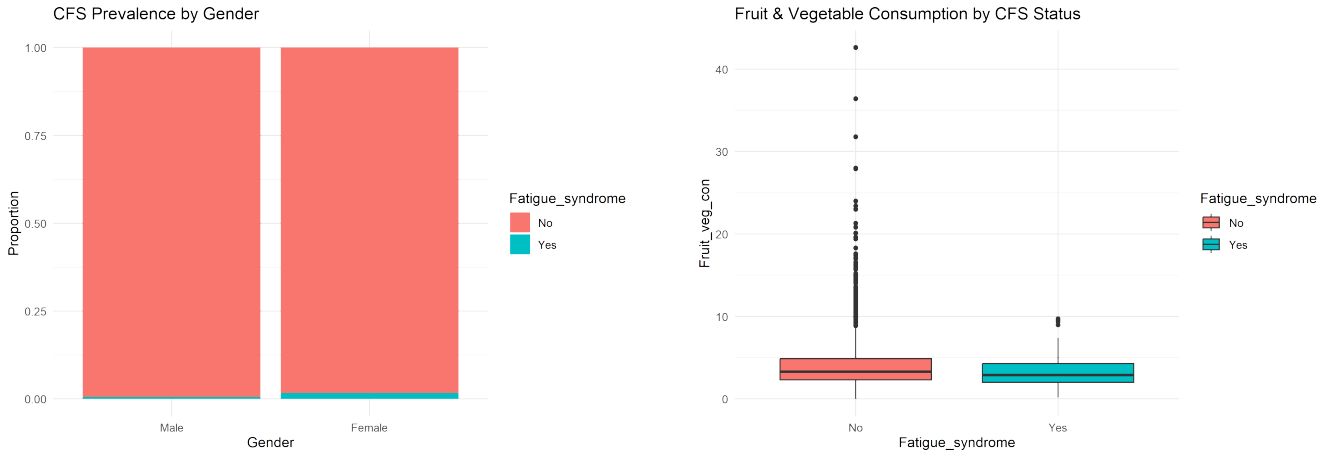


Figure 1: CFS prevalence by age group

Figure 1 shows the prevalence of CFS across different age groups. There is a clear class imbalance that is the case among all variables. We see an increase in the number of cases of CFS

as the age range increases. Figure 2 shows a larger proportion of CFS cases in females as opposed



(a) CFS prevalence by gender

(b) Fruit and vegetable consumption by CFS status

Figure 2: Demographic and behavioral characteristics associated with CFS

to males. The box plot in Figure 2 is showing fruit and vegetable consumption by CFS status. At a glance, it does not appear to show any significant association with CFS.

3 Methods

The data for this analysis is obtained from the "Healthcare Survey" dataset available on Kaggle [5]. The dataset contains anonymized responses from the Canadian Community Health Survey (2019) which includes a wide range of variables related to individual health, demographic background, and lifestyle habits. The variable `fatigue_syndrome` is used as the binary response variable. Logistic regression is particularly suitable for this analysis because the outcome variable is binary.

The logistic regression model takes the form:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where $P(Y=1)$ is the probability of having CFS, the parameterβ_0 represents the intercept and $\beta_1, \ldots, \beta_k$ are coefficients representing the log-odds of the predictors.

Based on our exploratory data analysis, literature review, and theoretical considerations, the variables included in the model can be categorized as demographic or clinical/behavioral, as summarized in Table 2.

To evaluate the adequacy of our logistic regression model, we assess model fit using likelihood-ratio tests, examine the significance of predictors via Wald tests, and we also perform analysis of residuals to identify influential observations.

Additionally, given the pronounced class imbalance (9353 No vs. 107 Yes), we implemented a weighted logistic regression model to up-weight the minority class during model fitting. We further

Category	Variable Description
Demographic Factors	
Age	Categorical: 18–34, 35–49, 50–64
Gender	Binary: Male, Female
Clinical and Behavioral Factors	
BMI Category	Underweight/Normal vs. Overweight/Obese
Diabetes	Binary: Yes, No
Anxiety Disorder	Binary: Yes, No
Sleep Apnea	Binary: Yes, No
Stress Level	Ordered factor from *Not stressful at all* to *Extremely stressful*
Mental Health State	Ordered factor from *Poor* to *Excellent*
Fruit and Vegetable Consumption	Continuous variable
Cannabis Use	Binary: Yes, No

Table 2: Summary of variables used in the logistic regression model

assess model performance using sensitivity and specificity measures, and ROC curve analysis to ensure that predictive performance for the minority class is addressed.

Given that our initial logistic regression model includes several predictors, some of which are not statistically significant, we refine the model to improve interpretability. Here, we applied both AIC-based and LASSO-based selection. For AIC, the model was reduced to a subset of variables that included age, gender, anxiety disorder, sleep apnea, stress level, mental health state, fruit and vegetable consumption, and cannabis use. This model had a slightly lower AIC than the full model.

The LASSO regression further reduced the number of variables by shrinking several coefficients to zero. It emphasized Age, Mental Health, Anxiety Disorder, Sleep Apnea, and Stress Level. Both approaches consistently selected Age, Anxiety Disorder, Sleep Apnea, Stress Level, and Mental Health as important predictors.

4 Results

To evaluate the overall model fit, we compared the full weighted logistic regression model to a null model using a likelihood ratio test. The test yielded a deviance difference of 1724.8 with 14 degrees of freedom, resulting in a highly significant p-value ($p < 2.2 \times 10^{-16}$). This indicates that the inclusion of predictors significantly improves model fit over the intercept-only model.

The statistical significance of individual coefficients was assessed using Wald tests, as reported in the model summary. Predictors such as age, gender, anxiety disorder, sleep apnea, stress level, and mental health state had p-values below 0.05, indicating strong evidence that these variables are associated with the odds of reporting CFS.

The model had a residual deviance of 5172.8 on 9445 degrees of freedom and an AIC of 5202.8, indicating good model fit relative to the null model.

Predictor	**Interpretation of Odds Ratio**
Age 35–49	3.26 times higher odds of reporting CFS compared to 18–34
Age 50–64	9.07 times higher odds of reporting CFS compared to 18–34
Female	2.96 times higher odds compared to males
Anxiety Disorder	2.51-fold increase in odds of CFS
Sleep Apnea	1.87 times higher odds
Stress (Linear Component)	Approximately 5.13 times higher odds with increasing stress
Mental Health (Lowest Tier)	Approximately 2.37 times higher odds with poorer mental health
Fruit/Veg Consumption	0.93 odds ratio per unit increase (protective effect)
Cannabis Use	1.47 times higher odds of reporting CFS

Table 3: Key findings from weighted logistic regression model: interpretation of odds ratios

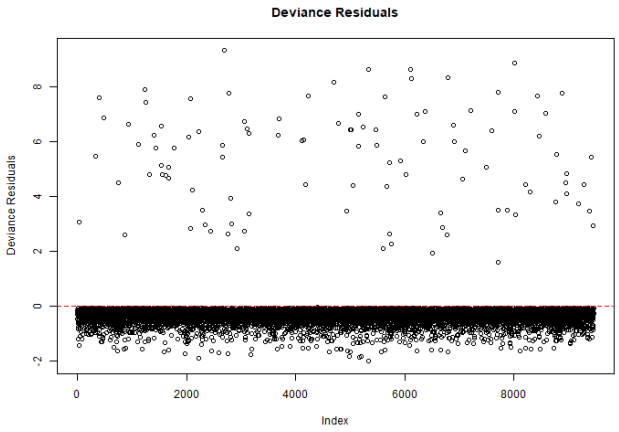
To further assess model fit and identify potential influential observations or outliers, we examined several residual diagnostics. Figure 3a shows the deviance residuals. Most residuals clustered around zero, indicating overall good model fit, though a few higher residuals suggest mild misfit for a small subset of observations.

Figure 3b plots standardized Pearson residuals against fitted values. The residuals generally follow a random scatter, which supports the assumption that the logistic link function is appropriate.

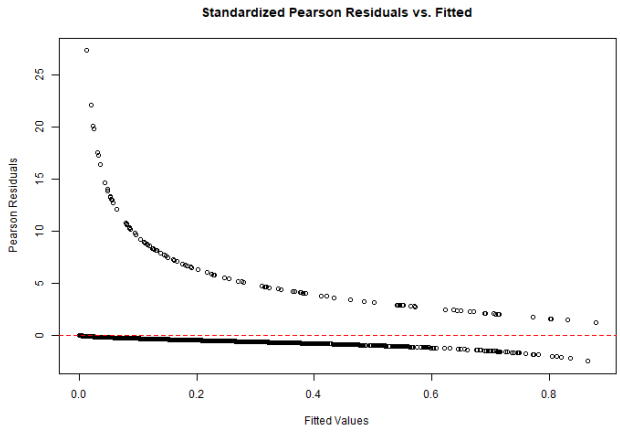
Cook's distance (Figure 4a) was used to assess influence. While most observations had low influence, a small number exceeded the conventional threshold of $4/n$, indicating they may disproportionately affect the estimated coefficients.

Lastly, Figure 4b visualizes the relationship between leverage and Cook's distance. A few observations combined high leverage and influence.

Together, these diagnostics suggest that the logistic regression model is generally well-behaved with good fit. However, a few potentially influential observations exist and could be examined more closely in the future.

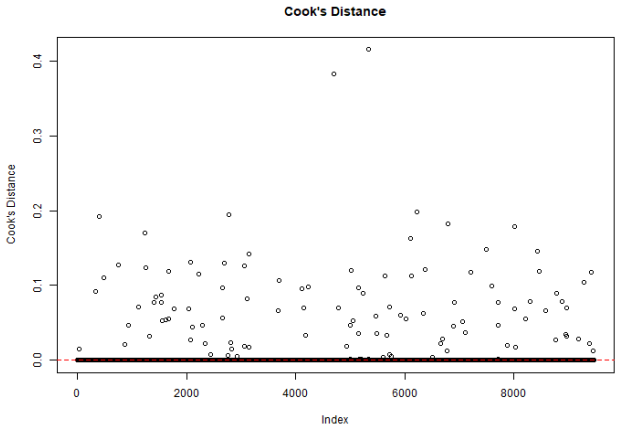


(a) Deviance residuals plot

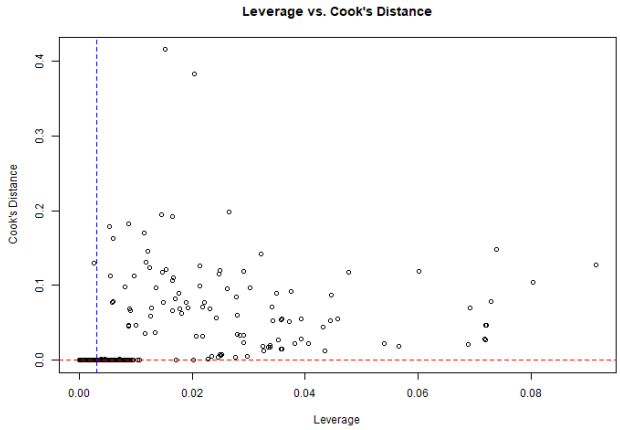


(b) Standardized Pearson residuals vs. fitted values

Figure 3: Residual diagnostics for logistic regression model



(a) Cook's distance plot for influential observations



(b) Leverage vs. Cook's distance plot

Figure 4: Influence diagnostics for logistic regression model

Classification Performance and ROC Analysis

To assess model performance in distinguishing individuals with and without CFS, we computed classification metrics and plotted the ROC curve.

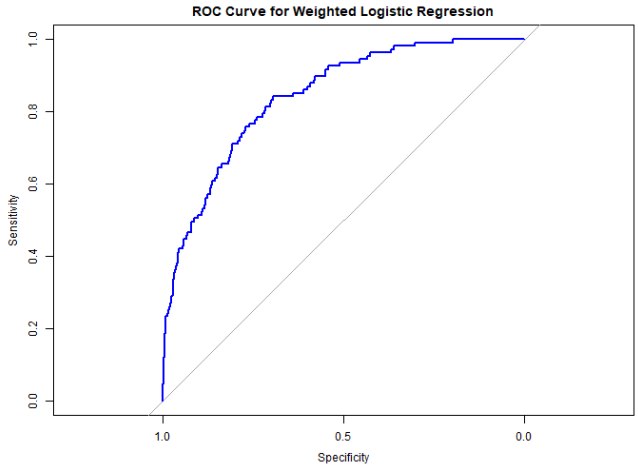


Figure 5: ROC Curve for Weighted Logistic Regression

The AUC of 0.842 indicated strong discriminative ability. The best threshold was found to be approximately 0.079, yielding improved specificity and balanced accuracy.

5 Discussion

Older age groups showed significantly higher odds of CFS, which is consistent with the hypothesis that the cumulative effect of physiological and psychological stressors over time may contribute to the development or worsening of fatigue symptoms. Females had substantially higher odds compared to males, reflecting the higher reported prevalence of CFS among women in epidemiological studies, potentially due to hormonal, immune, or psychosocial differences.

Psychological and sleep-related conditions like anxiety disorders, sleep apnea, higher perceived stress, and poor self-rated mental health, were strongly associated with increased odds of CFS. These results are intuitive, given that CFS is linked with dysregulation of the stress-response system, poor sleep quality, and mental health comorbidities.

Lifestyle factors such as fruit and vegetable consumption and cannabis use showed more modest associations. Lower intake of fruits and vegetables may be a marker of overall poorer diet quality and reduced intake of essential nutrients that support immune and metabolic function. Meanwhile, cannabis use may reflect self-medication for underlying symptoms such as pain, anxiety, or sleep disturbance, all of which are commonly reported among individuals with CFS. However, the causality of this relationship warrants further investigation.

The application of weighted logistic regression successfully addressed the challenge of severe class imbalance in our dataset. Key predictors identified consistently across unweighted, AIC-selected, and LASSO-regularized models were further reinforced in the weighted model. These included age, gender, anxiety disorder, sleep apnea, perceived stress, and mental health status. The role of lifestyle factors such as diet and cannabis use highlights potential avenues for public health interventions and further clinical study.

The ROC curve and confusion matrix confirmed good classification performance overall. Threshold tuning using Youden's J statistic [7] enabled a better balance between sensitivity (84.1%) and specificity (69.5%), improving balanced accuracy to 76.8%.

Future analyses could further investigate the trade-offs between sensitivity and specificity by exploring a range of thresholds and cost-sensitive evaluation frameworks. Within the context of generalized linear models, the weighted logistic regression framework with tuned thresholds provides an interpretable and effective approach to modeling rare outcomes like CFS.

6 Conclusion

This analysis investigated the predictors of Chronic Fatigue Syndrome (CFS) using logistic regression on data from the 2019 Canadian Community Health Survey. Given the rarity of CFS in the sample, special attention was paid to class imbalance, leading to the use of a weighted logistic regression approach. Through variable selection and threshold optimization we identified several key factors associated with increased or decreased odds of reporting CFS.

Demographic variables such as age and gender were significantly associated with CFS, with older adults and females at increased risk. Clinical and psychological factors—including anxiety disorders, sleep apnea, perceived stress, and poorer mental health also showed strong positive associations.

Model diagnostics confirmed that the logistic regression model fit the data well, with few influential points and no major violations of modeling assumptions. By selecting an optimal threshold using Youden's J statistic, we improved the balanced accuracy to 76.8%, further enhancing the model's practical utility. This investigation provides insights that could support further clinical and public health research on CFS.

References

[1] Hickie, I., Davenport, T., Wakefield, D., et al. (2006). Post-infective and chronic fatigue syndromes precipitated by viral and non-viral pathogens: prospective cohort study. *BMJ*, 333(7568), 575.

[2] Jason, L. A., Richman, J. A., Rademaker, A. W., et al. (2001). A community-based study of chronic fatigue syndrome. *Archives of Internal Medicine*, 159(18), 2129–2137.

[3] Nisenbaum, R., Reyes, M., Mawle, A. C., et al. (2004). Prevalence of chronic fatigue syndrome in a population-based sample. *American Journal of Epidemiology*, 160(3), 245–253.

[4] White, A. T., Light, A. R., Hughen, R. W., et al. (2009). Identification of potential biomarkers for chronic fatigue syndrome using gene expression profiling. *Pharmacogenomics*, 10(2), 243–252.

[5] Hirapara, A. (2022). Healthcare Survey. Retrieved from `https://www.kaggle.com/datasets/aradhanahirapara/healthcare-survey`

[6] Victora, C.G., et al. (1997). The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. *International Journal of Epidemiology*, 26(1), 224–227.

[7] W. J. Youden. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.