Effect of Socio-Economic and Lifestyle Factors on Depression

Rudraharsh Tewary

The following study analyzes the Effect of Socio-Economic Factors like Race and Education coupled with Lifestyle Factors on the odds of someone suffering from depression. Bayesian analysis was carried on Data sourced from National Health Interview Survey (NHIS) carried out by IPUMS. Analysis was done using a hierarchical logistic regression model. Results show us that Native-Americans are a vulnerable minority in terms of suffering from depression and that lifestyle factors exhibit a strong influence on the log odds(+/- 0.7) of someone being depressed.

Introduction

The primary motivation for this project is to analyze whether social and economic factors like Race and Education impact the odds of someone being depressed. To perform this analysis, health survey data from Integrated Public Use Microdata Series (IPUMS) was used, pertaining to the year 2022. The aim is to assess how different levels of education and a person's race, in conjunction with their lifestyle choices affect the outcome of depression, as well as to discern how strong said link is.

Previous work has shown that minority groups are more predisposed to showcasing symptoms of Depression (Dunlop DD 2003, Bailey RK 2019). Their studies showed that minority groups would suffer from higher rates of depression due to discrimination and lesser avenues for success as compared to caucasians.

While work done by (Sarris J. 2020) showcased that lifestlye factors like consumption of tobacco, screen-time, healthy diet etc. have a noticeable impact on the log-odds of someone exhibiting Major Depressive Disorder (MDD) with people showcasing healthier lifestyle choices and no-smoking exhibiting the lowest odds of Depression.

Also, research done by (Cohen AK 2020, McFarland MJ 2015, Lingli Li 2022) showcases that successful attainment of higher levels of education. That is, education level of a bachelor's degree or above corresponds with inverse odds of depression.

Data

The Data being used for the study is National Health Interview Survey Data (NHIS) collected by IPUMS (Lynn Blewett 2023) in the year 2022. The inclusion criteria for respondents in our study was for them to be (1) classified as legal adults (Age $\geq$ 18); (2) To not have responded with "unknown" in any item on the survey questionnaire. Because of this, the original data sample size (n = 35,115) was subsetted to (n = 7,128) respondents. Aforementioned Data only surveyed citizens of the United States.

Missing Values

The Dataset does not contain any missing values. However, there is a substantial portion of respondents who answered "Unknown"/"No answer" on parts of the questionnaire, those data points have been removed leaving us with a complete dataset.

Selected Survey Data

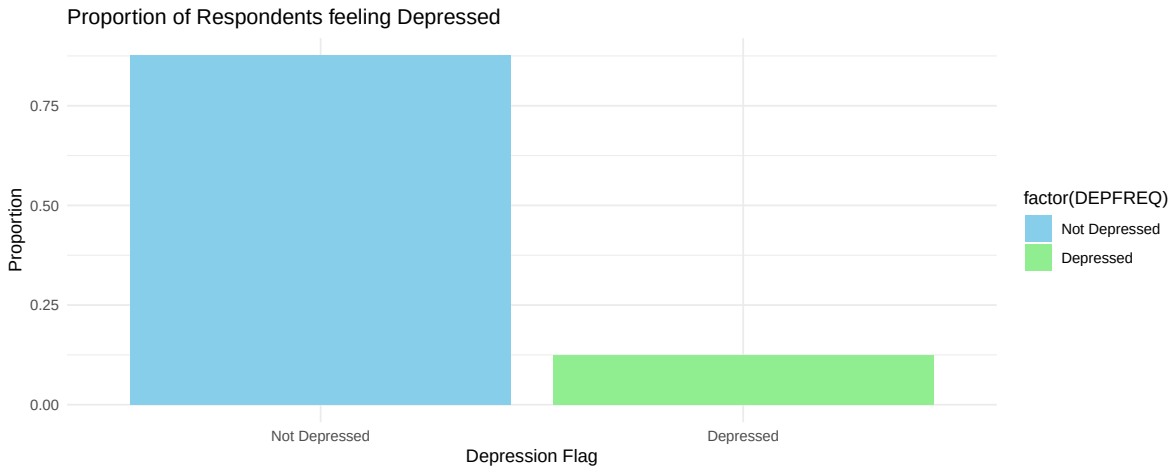
The following variables were selected post data pre-processing and Exploratory Data Analysis (EDA). 'HRSLEEP', an integer variable, this tells us the average hours of sleep a respondent gets with a range from 0 to 24. 'SMOKESTATUS2', a categorical variable, it is divided into multiple categories showing us how frequently the respondent smokes, with the levels being never smoked, Former smoker, Current smoker, frequent smoker, daily smoker. 'MOD10FTP', this details the amount of moderate physical activity greater than 10 minutes performed by a respondent with it having the following levels, unable to do it, never done, done yearly, done monthly, done weekly, done daily, extreme frequency.

The following variables were of special interest.'EDUC', which states the education level of a participant, ranging from a high school diplotma to a doctorate degree and 'HISPRACE' which tell us the race of the respondent are special points of interest and will be modeled hierarchically, these variables were pre-selected since the beginning.

Finally, the response variable is 'DEPFREQ', which is a binary variable that tells us whether a respondent is depressed or not. It is a binary variable with 1 standing for Depressed and 0 standing for not depressed.

Variable Trends and Associations

We begin by looking at the proportion of depressed respondents in the dataset:



Around 12.5% of respondents in the dataset are depressed. On the basis of that, we now proceed to evaluate trends and relations of our selected variables.

We proceed to showcase proportions of depression by race:

```
# A tibble: 5 x 2
  proportion race_label
       <dbl> <chr>
1     0.104  Hispanic
2     0.136  White
3     0.1    Black
4     0.0664 Asian
5     0.203  Native American
```

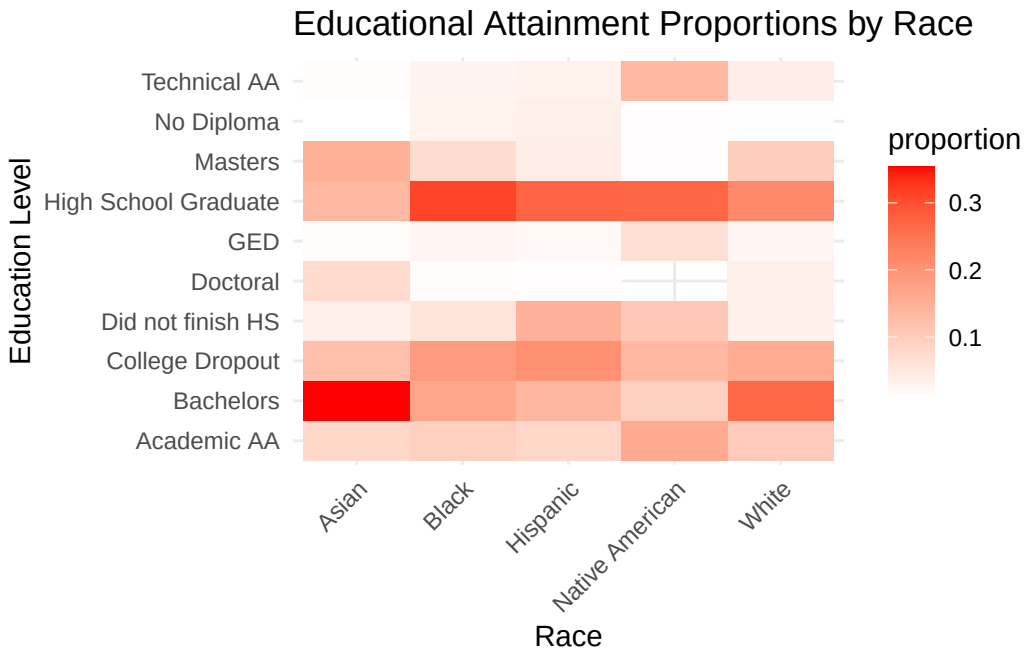
We notice that Asians seem to have the lowest proportion of Depression in the survey, while Native Americans report the highest proportion of Depression by far, at around 20%.

We then proceed to analyze proportions of Depression by education level:

```
# A tibble: 10 x 2
   proportion educ_label
        <dbl> <chr>
 1     0.157  Did not finish HS
 2     0.127  No Diploma
 3     0.132  High School Graduate
 4     0.178  GED
 5     0.168  College Dropout
 6     0.136  Technical AA
 7     0.123  Academic AA
```

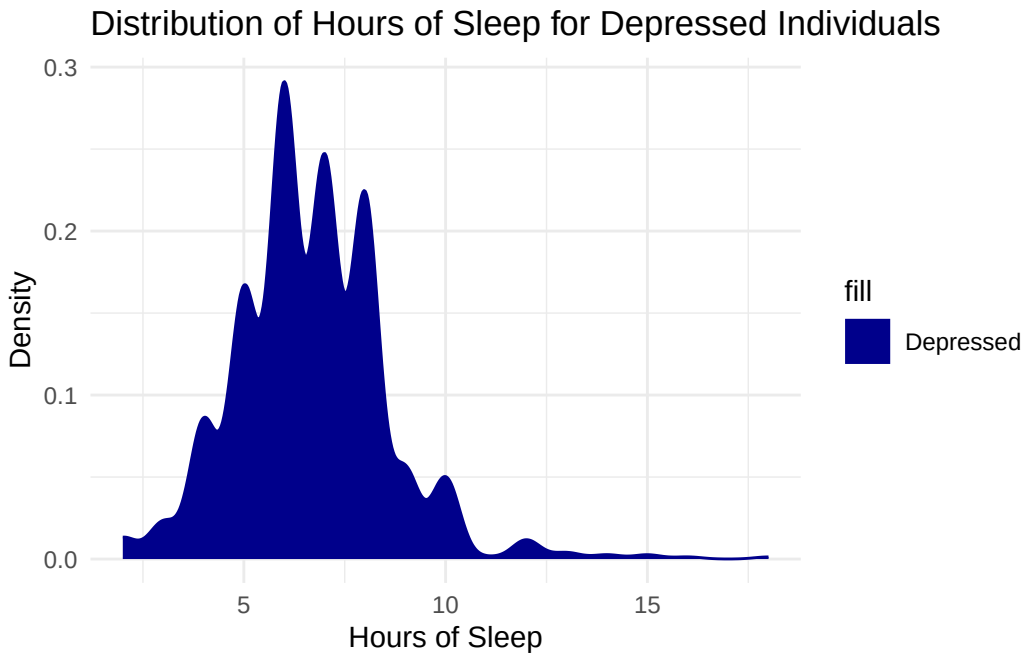
```
 8     0.0859 Bachelors
 9     0.0843 Masters
10     0.109  Doctoral
```

Here, we see that higher concentrations of depression are found in people who finished their education at a level of less than an AA degree. With respondents at the level of 'College Dropout' and below making a Majority chunk of Depressed respondents.



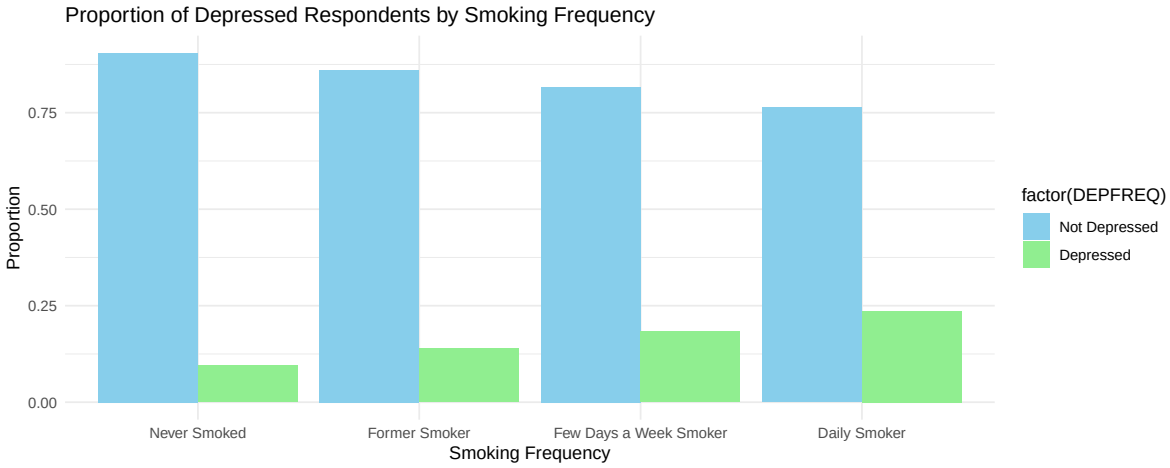
Here we notice that Asians, followed by White people have the highest proportion of people who have advanced educational attainments like a bachelor's degree or above. While other minorities showcase high proportions at the levels of High School Graduate and/or College Dropout. Of special interest are Native-Americans who show very low presence at higher education levels like Master's and Doctoral Degrees, though it has to be mentioned that they also account for a very small portion of the dataset (1%).

We now proceed to analyze the relation between sleep and depression:



According to the data, depressed individuals in general have much more variability in terms of sleep duration reported, and have greater portion of people sleeping above 8 hours or less than 5 hours. While in terms of people who reported not feeling depressed, their sleep duration is contained in the 6-8 hour region.

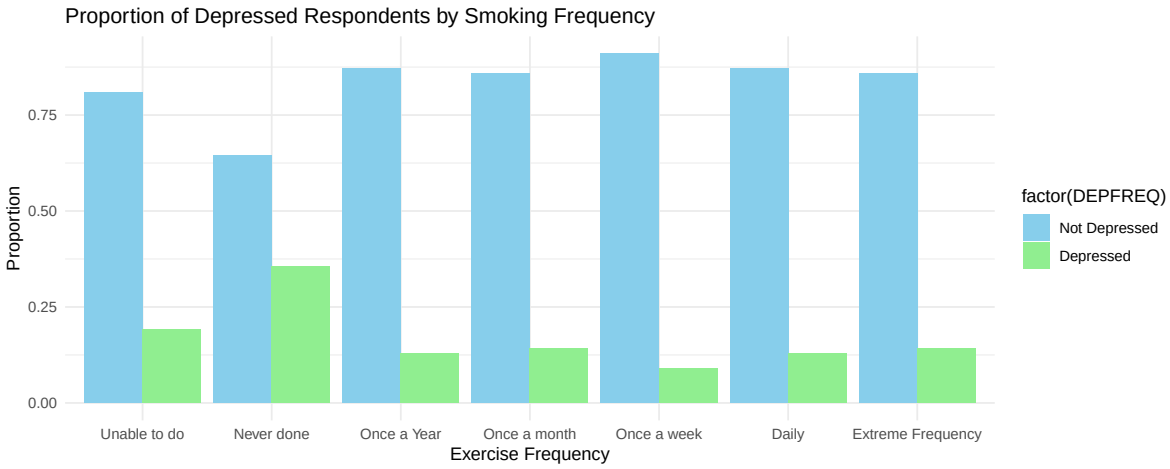
Now, concerning frequency of Smoking and Depression, we see the following trend:



Here there is a noticeable trend where the groups which reported smoking at a higher frequency, the proportion that reported feeling depressed increased and the proportion that reported the

opposite decreased. That is, higher frequencies of smoking correspond with higher chances of being depressed.

We now analyze the association between moderate physical activity and depression:



Here we see an inverse association, wherein groups which reported performing moderate physical activity at a higher frequency had a greater proportion of people who reported not feeling depressed.

Methods

As the response variable (Depression outcome), is represented as a binary variable, we will model it with a logistic regression :

$$y_i|\pi_i \sim Bernoulli(\pi_i)$$

Where y is the response variable 'DEPFREQ' and π is the specified model.

For the chosen variables of our model, we label them as follows:

1. Race : this covariate represents the race of a given respondent.
2. Sleep : this covariate represents the daily average hours of sleep of a respondent.
3. Smoke : this covariate represents the smoking frequency of a respondent. As this is a categorical variable, the baseline would be someone who never smoked.
4. Modex : this covariate represents how frequently a respondent performs moderate physical activity over a duration of 10 minutes. As this is a categorical variable, the baseline for this would be someone who is Unable to exercise.
5. Educ : this covariate represents the highest education level of a given respondent.

Now, due to the trends we observed in the data section, we will define two models for π, One which would model effects of race and education hierarchically and which would not have any hierarchy.

As the desire is for the data to shape the model, all selected priors would be weakly informative, following the $N(0,1)$ distribution, we then define the priors as:

$$\alpha \sim N(0,1)$$

$$\beta_e^{educ} \sim N(0, \sigma_{educ}^2)\ Where\ e = 1,2,3,4,5,6,7,8,9,10$$

$$\beta_r^{race} \sim N(0, \sigma_{race}^2)\ Where\ r = 1,2,3,4,5$$

$$\beta_{sleep} \sim N(0,1)$$

$$\beta_{modex} \sim N(0,1)$$

$$\beta_{smoke} \sim N(0,1)$$

Where, σ_{educ} and σ_{race} are defined as:

$$\sigma_{educ} \sim N(0,1)^+$$

$$\sigma_{race} \sim N(0,1)^+$$

We then proceed to define both models, first being the normal non-hierarchical model represented as:

$$\pi_i = logit^{-1}(\alpha + \beta_{educ} * educ_i + \beta_{race} * race_i + \beta_{sleep} * sleep_i + \beta_{modex} * modex_i + \beta_{smoke} * smoke_i)$$

And the second one being the hierarchical model represented as:

$$\pi_i = logit^{-1}(\alpha + \beta_{e[i]}^{educ} + \beta_{r[i]}^{race} + \beta_{sleep} * sleep + \beta_{modex} * modex + \beta_{smoke} * smoke)$$

The model will be fit using R stan, where sampling will be carried out through Markov Chain Monte Carlo (MCMC), the specific algorithm that will be used is No U-Turn Sampling (NUTS).

For the purposes of model validation, we will use a traceplot to evaluate model convergence, and then posterior predictive checks to evaluate if the model has accurately captured trends in the data. Both models have been fitted with 4 chains at 2000 iterations.

Alternative models tried

I also attempted to fit a hierarchical model with an interaction term between race and education (as the data suggests both of these covariates are co-dependent) the model was defined as:

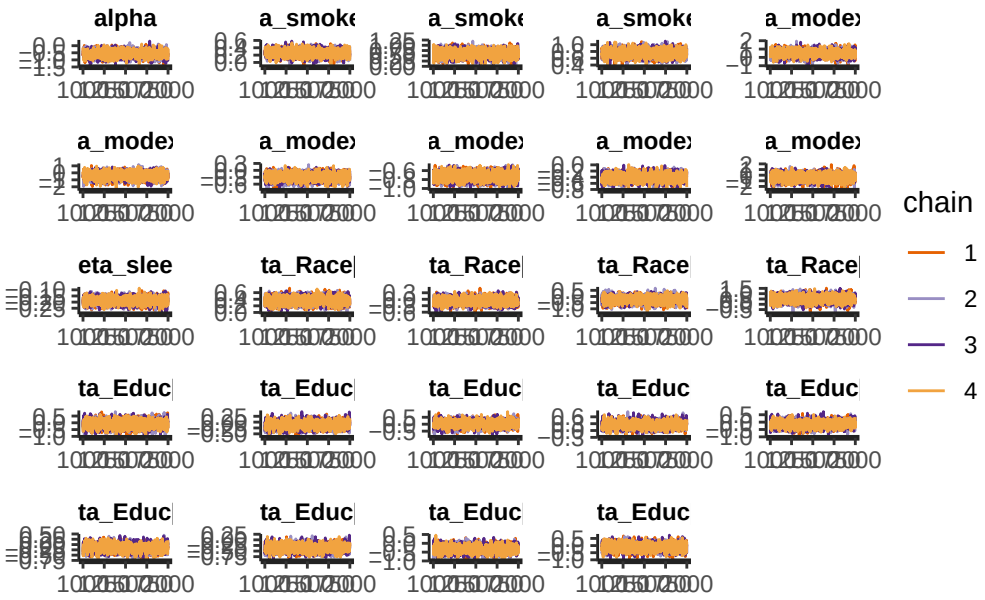
$$\pi_i = logit^{-1}(\alpha + \beta^{educ}_{e[i]} + \beta^{race}_{r[i]} + \beta_{sleep} * sleep + \beta_{modex} * modex + \beta_{smoke} * smoke + \beta^{educ\_race}_{e[i],r[i]})$$

However, I was unable to use R to perform any inference on this model, as anytime a function would be called to evaluate the fitted object, R would crash.

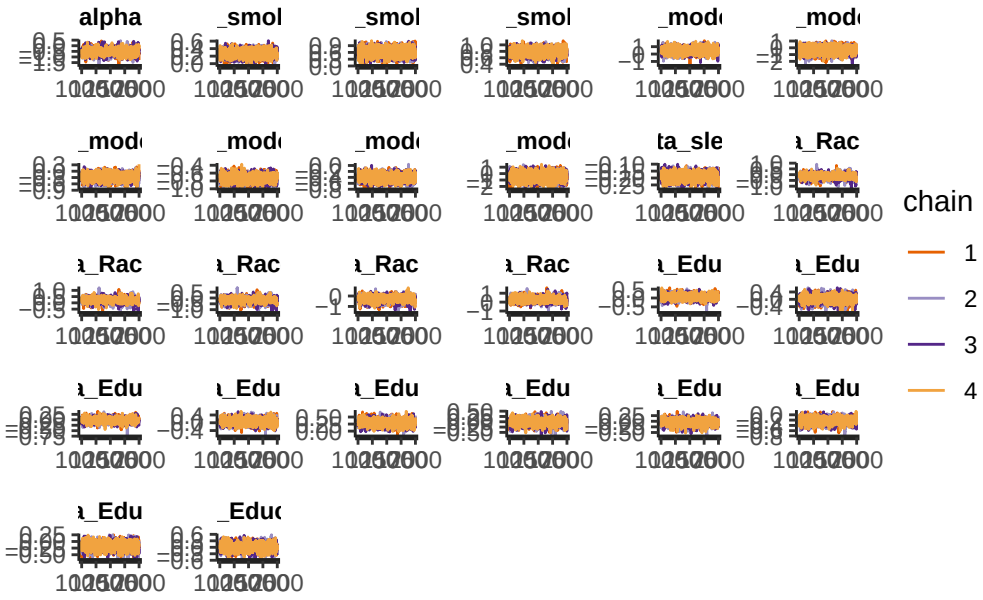
Results

Model Convergence and Selection

We proceed to first evaluate the convergence of both models using trace plots: We first analyze the non-hierarchical model:



We see that all the parameters seem to have converged well, we proceed to do the same for the Hierarchical model next:



We seem to have convergence for this model as well. Now, we will proceed to compare the performance of both models based on Leave-One-Out Cross-Validaton (LOO-CV).

Model	elpd_diff	se_diff
Hierarchical Model	0.0	0.0
Non-Hierarchical Model	-2.1	1.5

Upon comparing the Expected log Pointwise Predictive Density (ELPD) of both models, the Hierarchical model seems to be better.

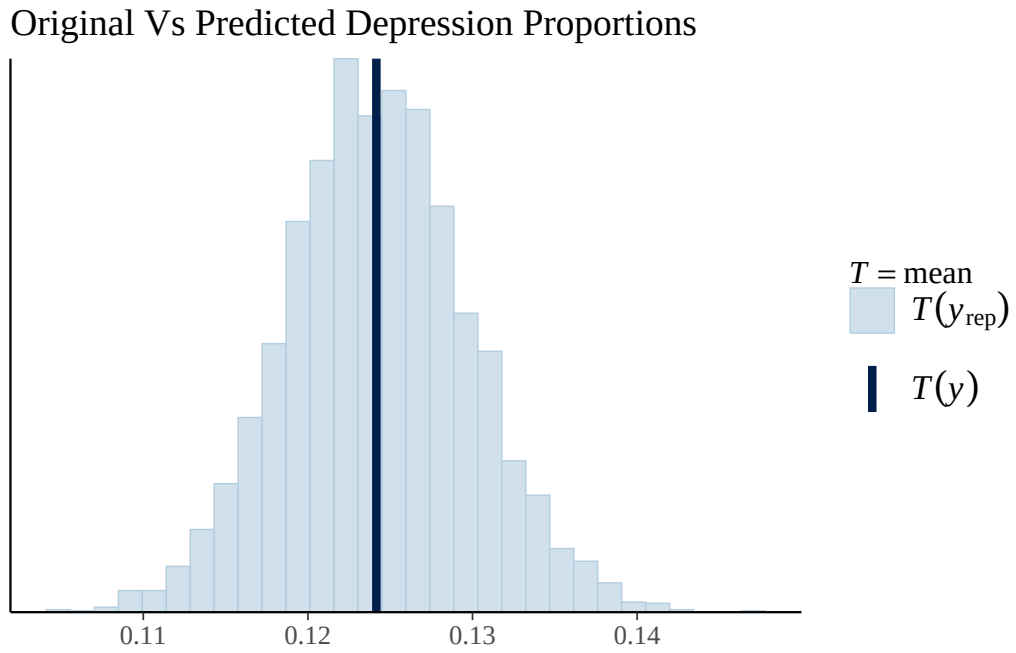
However, when we look at the LOO computations for both models, we notice the following:

Model	elpd_loo	p_loo	looic
Hierarchical Model	-2541.3(±53.9)	20.3(±0.8)	5082.7(±107.8)
Non-Hierarchical Model	-2543.4(±54.0)	23.6(±0.9)	5086.9(±108.0)

It appears that both models have a relatively large standard error in their ELPD estimates. Therefore, we cannot say with complete confidence that the hierarchical model is outright better. However, based on the previous analysis and the effects of race and education on the response we will choose to base our analysis and results on the Hierarchical model.

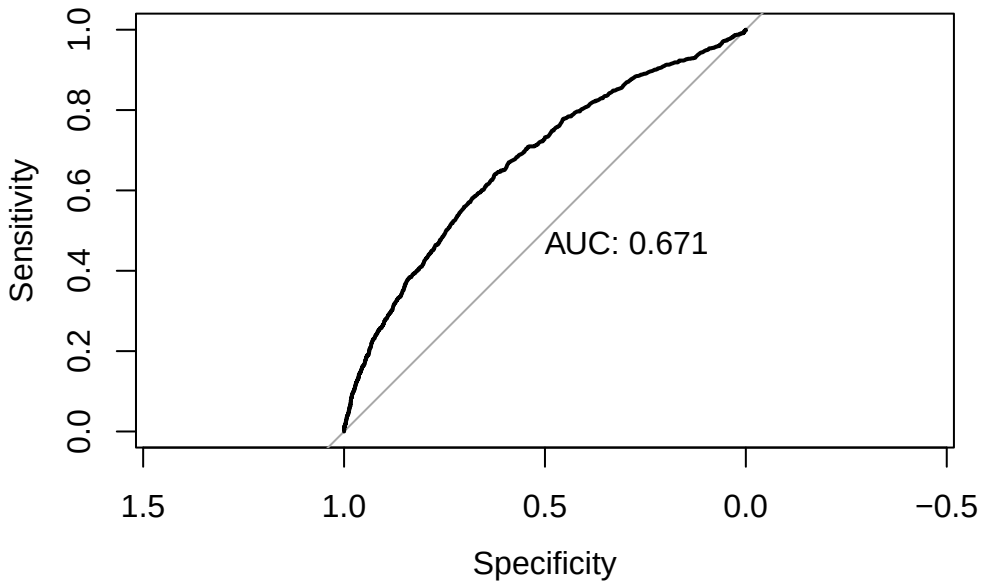
Posterior Predictive Checks

We can now begin by performing some posterior predictive checks on our model:



The plot suggests that posterior replications are able to match the proportion of positive outcomes in the true Dataset at 12.5%.

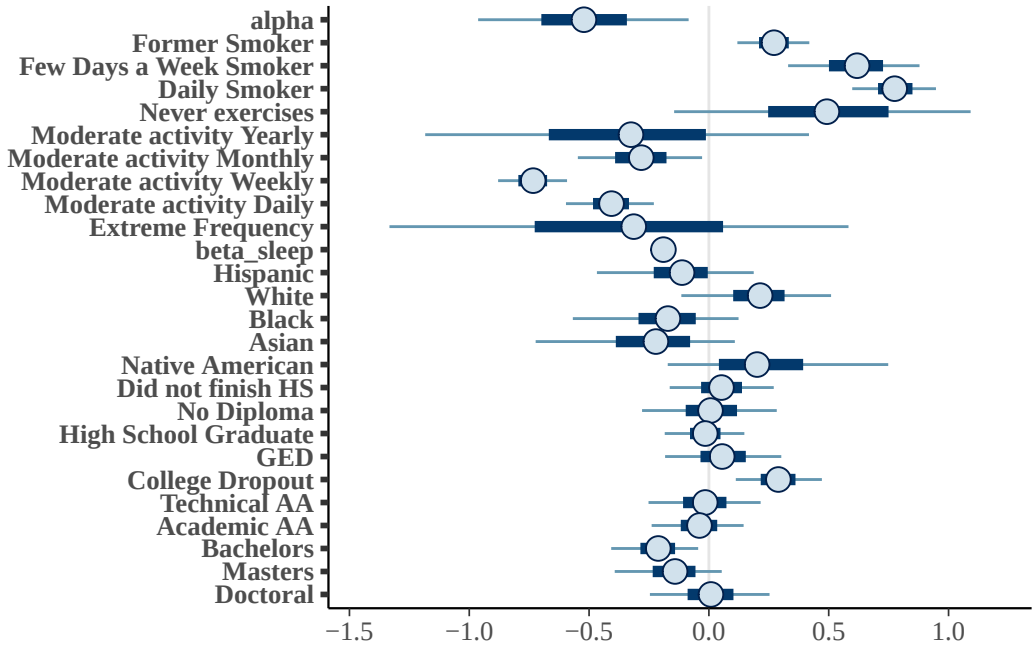
We then proceed to check the performance of our model by looking at its Reciever Operating Characteristic (ROC) curve:



Unfortunately our model is not able to generalize well, as the Area Under the Curve (AUC) is around 0.671, which is only greater than the random guessing threshold of 0.5 by 0.171, which is only 17.1% better than random guessing. This shows that our model does not have great prediction accuracy.

Parameter Estimates and Analysis

So, we now proceed to get the Parameter estimates from the Hierarchical model:



First, analyzing the lifestyle factors, we see more frequent levels of smoking correspond with a higher positive coefficient for depression according to the model. On the other hand, higher exercise levels and sleep tend to suggest the opposite, with the greatest boosts against depression occuring the class of people who showcased moderate physical activity weekly according to the model, one thing that would stand out though is the wide credibility intervals on the levels of 'Never exercises', 'Moderate activity Yearly' and 'Extreme Frequency' this is because the dataset contains a low population of respondents who reported physical activity at these levels. Finally, we see that hours of sleep have a negative correlation with depression, suggesting that if someone gets the optimal amount of sleep/ more sleep their odds of feeling depressed lower significantly, not to mention this specific coefficient also have very narrow 50% and 90% credibility intervals

Now, focusing on the effects of race, the model suggests that someone who is White/Native American has higher likelihood of being depressed. While people from other races show a lower likelihood of depression. However, for all races we seem to notice wider bounds on the 50% and 90% Credibility intervals.

In terms of education, we see a decreasing trend overall, that is, a higher level of education as compared to the previous one would usually result in a lower coefficient for depression, with the effect becoming negative when we get to the levels of 'Bachelor's' and 'Master's' Degrees. The only unique scenario here would be 'Doctoral' Degrees, where we see that the coefficient suggests that the likelihood of being depressed is more than the other two higher education levels and also higher than average.

Now, we proceed to look at the coefficients:

Coefficient	Mean ± SE Mean	R_hat
alpha	-0.52 ± 0.01	1
Former Smoker	0.27 ± 0.00	1
Few Days a Week Smoker	0.61 ± 0.00	1
Daily Smoker	0.78 ± 0.00	1
Never exercises	0.49 ± 0.01	1
Moderate activity Yearly	-0.35 ± 0.01	1
Moderate activity Monthly	-0.29 ± 0.00	1
Moderate activity Weekly	-0.73 ± 0.00	1
Moderate activity Daily	-0.41 ± 0.00	1
Extreme Frequency	-0.34 ± 0.01	1
beta_sleep	-0.19 ± 0.00	1
Hispanic	-0.12 ± 0.01	1
White	0.21 ± 0.01	1
Black	-0.18 ± 0.01	1
Asian	-0.25 ± 0.01	1
Native American	0.23 ± 0.01	1
Did not finish HS	0.05 ± 0.00	1
No Diploma	0.01 ± 0.00	1
High School Graduate	-0.02 ± 0.00	1
GED	0.06 ± 0.00	1
College Dropout	0.29 ± 0.00	1
Technical AA	-0.02 ± 0.00	1
Academic AA	-0.04 ± 0.00	1
Bachelors	-0.22 ± 0.00	1
Masters	-0.15 ± 0.00	1
Doctoral	0.01 ± 0.00	1

Looking at the coefficients, we see that in terms of lifestyle factors, Daily smoking has the strongest impact on feeling depressed, so if someone claims to be a daily smoker, the model suggests that their log-odds of feeling depressed increase by 0.78. Whereas, for the negative correlation we see that moderate physical activity weekly corresponds with a log odds decrease of 0.78.

Focusing on the terms modelled hierarchically. We see that being Native-American corresponds to an increase of 0.23 units in the log-odds of being depressed, closely followed by being White having a log-odds increase of 0.21 units. Now, looking at education levels, it appears that across the board lower education levels correspond to an increase in the log-odds of feeling depressed. We see that people who did not finish High School, No diploma and College Dropouts woulds cause a log-odds increase in feeling depressed. Though of note are college dropouts, showcasing that they are the most susceptible to feeling depressed with a log-odds increase of 0.29. In terms of negative correlation, we see that people who finished higher education and got a

Bachelor's or a Master's showcase a negative relation with depression, with the highest being Bachelor's degree holders at a log-odds decrease of 0.22.

(Note:- All discussions of impact of coefficients on the log-odds of feeling depressed assumes that the other coefficients are held constant.)

Discussion

Looking at the results, we find that depression is affected by a multitude of lifestyle and socioeconomic factors. Focusing on the socioeconomic factors, we find that people who are 'White' and 'Native-American' have a positive relation with depression in terms of log-odds. Focusing on Native-Americans, we can claim that this result is not that surprising considering the fact that we also have an established relationship between education levels and depression, and we have seen that according to the data, Native-Americans have the lowest representation at higher education levels like Bachelor's,Master's and Doctorate Degrees. This could indicate that there are other factors like economic inequality and discrimination that are affecting their rise to a better socio-economic status. On the other hand, we also see that the data suggests being White suggests higher log-odds of depression this could possibly be due to the fact that White people might be over-represented in the dataset accounting for 66% of all respondents in the subset which might lead to inflation in their log-odds of depression, as according to the previous analysis we see that White people seem to have good-representations in higher-education levels and overall are more socio-economically successfuly than most other minority groups.

However, our model also suggests that lifestyle factors have a stronger impact on the log-odds of depression. We see that daily smoking showcases the maximum possible increase in the log-odds of a depression outcome at +0.7. While on the other end moderate physical activity shows the opposite trend where performing physical activity weekly will decrease the log-odds of someone being depressed by 0.7. And sleep has a fixed relationship where more sleep corresponds with lower log-odds in depression.

Therefore, we can draw a mixed result where race and education level do have a bearing on feelings of depression and that certain categories in both sections are more susceptible to feelings of depression. However, we also see that simple lifestyle choices like smoking and exercise play a crucial role in this scenario, suggesting that people suffering from depression who stop smoking, exercise more and get better sleep might see an improvement in their mental health.

Though, there are definitely certain weak points with the analysis. First one being that the model itself does not have a good ROC-AUC , telling us that its predictive powers are very weak thus any inference done from the model would be quite weak. Second being that the dataset is imbalanced in terms of Race, with white people making up 66% of the dataset, and Hispanic, Black, Asian and Native-Americans only occupying $16.7\%, 11.7\%, 4\%, 1.03\%$ of the data making any conclusions drawn on them significantly weak. Though of note is the fact

that even at such low-representation, Native-Americans have a startling proportion of people who exhibit depression.

For future work, the following can be avenues of interest. A bigger model with more parameters like income and disability status could have better predictive and inference capabilities. More data which might be able to tackle the under-representation of minorities would also significantly improve our ability to draw conclusions. Another interesting addition that can be made is to use global data instead of just data from the United States (As NHIS only surveys U.S. citizens), as one would notice we saw that higher education is negatively correlated with depression, but in the United States due to the costs of a degree being prohibitively expensive for the average person, higher education might also be correlated to higher income/wealth background for those who get it.

References

McFarland MJ, Wagner BG. Does a college education reduce depressive symptoms in American young adults? Soc Sci Med. 2015 Dec;146:75-84. doi: 10.1016/j.socscimed.2015.09.029. Epub 2015 Sep 28. PMID: 26513116; PMCID: PMC4676078.

Dunlop DD, Song J, Lyons JS, Manheim LM, Chang RW. Racial/ethnic differences in rates of depression among preretirement adults. Am J Public Health. 2003 Nov;93(11):1945-52. doi: 10.2105/ajph.93.11.1945. PMID: 14600071; PMCID: PMC1199525.

Bailey RK, Mokonogho J, Kumar A. Racial and ethnic differences in depression: current perspectives. Neuropsychiatr Dis Treat. 2019 Feb 22;15:603-609. doi: 10.2147/NDT.S128584. PMID: 30863081; PMCID: PMC6390869.

Sarris, J., Thomson, R., Hargraves, F. et al. Multiple lifestyle factors and depressed mood: a cross-sectional and longitudinal analysis of the UK Biobank (N=84,860). BMC Med 18, 354 (2020). https://doi.org/10.1186/s12916-020-01813-5

Cohen AK, Nussbaum J, Weintraub MLR, Nichols CR, Yen IH. Association of Adult Depression With Educational Attainment, Aspirations, and Expectations. Prev Chronic Dis 2020;17:200098. DOI: http://dx.doi.org/10.5888/pcd17.200098external icon.

Lingli Li, Wang Sun, Jinglan Luo, Hao Huang, Associations between education levels and prevalence of depressive symptoms: NHANES (2005–2018), Journal of Affective Disorders, Volume 301, 2022,Pages 360-367,ISSN 0165-0327, https://doi.org/10.1016/j.jad.2022.01.010, (https://www.sciencedirect.com/science/article/pii/S0165032722000155)

Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Annie Chen, Stephanie Richards, and Michael Westberry. IPUMS Health Surveys: National Health Interview Survey, Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D070.V7.3