An Analysis of Risk Factors Associated with Depression

Amy Stephen
University of Massachusetts Amherst
astephen@umass.edu

Sam Barber
University of Massachusetts Amherst
sbarber@umass.edu

***Abstract*—Depression affects more than 21 million adults in the US every year. Research has shown that depression is as a result of a complex interaction of factors, and in this project we set out to understand some of those risk factors. Utilizing data from NHANES, we fit Linear and Logistic Regression models and select the best-performing model. We analyze coefficients and covariates from the selected model to identify risk factors, and support our analysis with Directed Acyclic Graphs (DAGs).We found *Physical activity*, *BMI*, *Age*, and *Education Level* to be the most statistically significant predictors in classifying depression with our model.**

INTRODUCTION

According to a study conducted by the Center for Disease Control and Prevention, about 20 million adults in the U.S (about 8.4% of the U.S population) suffered from at least one major depressive episode [1]. Research suggests that depression occurs as a result of a combination of factors including but not limited to chemical/hormonal imbalances, genetic factors, lifestyle, and stressful life events.

Depression affects a person's mood, productivity, interpersonal relationships, self-perception, and in severe cases can lead to suicide. With these statistics and risks, it is important to understand the risk factors associated with depression as this is essential to identifying intervention strategies.

In this project, we use Linear and Logistic Regression Models to identify covariates associated with increased risk of depression in adults over 18 years of age.

I. DATA

Data for this project is taken from the 2017-2018 NHANES Database [4]. Before Data pre-processing, there were 5533 survey respondents. The outcome variable is obtained from responses to NHANES Depression Screening survey. The survey comprises of ten questions numbered DPQ010-DPQ100. All possible question values are presented in Table 1 below.

TABLE I
DEPRESSION SCREENING QUESTIONNAIRE VALUES

Code or Value	Value Description
0	Not at all
1	several days
2	more than half the days
3	nearly every day
7	refused
9	don't know

To compute the outcome variable, responses to the aforementioned screening questionnaire were summed up, with values ranging from $0 - 28$. For the Logistic Regression model, the outcome variable was binarized; scores less than 10 were categorized as *not showing depressive symptoms* and represented with a 0, scores greater than or equal to were categorized as *showing depressive symptoms* and were represented with a 1.

A summary of covariates used in the analysis is represented in the figure below.

Variable	Description	Datatype	% NA
RIAGENDR	Gender	Categorical	0
DMDEDUC2	Education level - Adults 20+	Discrete	4.8%
INDFMIN2	Annual family income	Categorical	4.6%
FIAPROXY	Proxy used during family interview	Categorical	4.6%
FSD032A	HH Worried run out of food	Categorical	4.5%
HUQ090	Seen mental health professional/past yr	Categorical	0
SLQ050	Ever told doctor had trouble sleeping?	Categorical	0
SLQ120	How often feel overly sleepy during day?	Discrete	0

Fig. 1. Coviariates Summary

For our analysis, we filtered for participants over eighteen years old, and dropped columns with percentage of missing values greater than 10%. We imputed the remaining missing values based on the mean value of the column.

II. EXPLORATORY DATA ANALYSIS

To better understand our data, we performed some exploratory data analyses.

A. Demographics

The NHANES survey design utilizes oversampling techniques to account for underrepresented groups in the population.

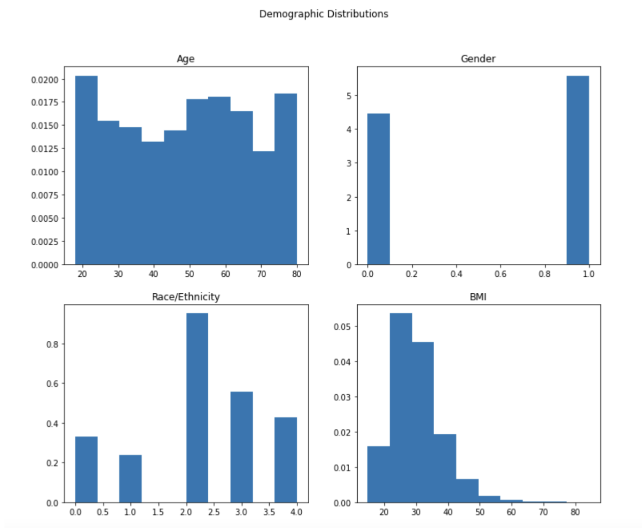


Fig. 2. Demographic Distributions

B. Outcome Variable

As shown in the figure 2, the dataset is highly unbalanced in relation to the outcome variable. Although this is in-line with our expectations considering the depression statistics in the U.S, it presents an issue when attempting to model the data. Due to the nature of regression models, they will tend to under-predict the minority class (in this case, showing depressive symptoms) which can lead to subpar model performance. We discuss this in the section on Model Selection.

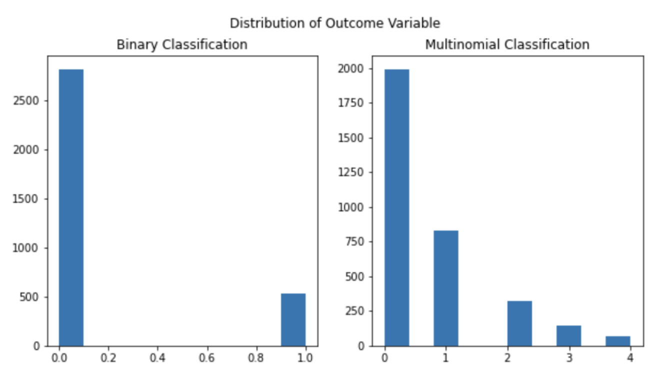


Fig. 3. Outcome Variable

C. Correlations between Predictors

When selecting features to be included in our model, we must first analyse the relationship between potential covariates. Highly correlated features present in the model may introduce bias and decrease model performance. The correlation heatmap in Figure 3 displays correlations between covariates with an absolute value greater than or equal to 0.1.

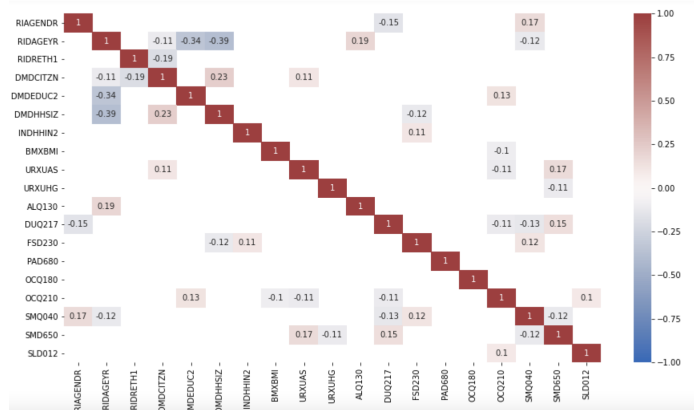


Fig. 4. Correlation Heatmap

Methods

III. Feature Selection

We decided to use feature selection to limit the number of features used in our models. Three main methods of feature selection were performed.

1) Forward Selection (Linear Model) - Greedy selection of best features.
2) Backward Elimination (Linear Model) - Greed elimination of least important features.
3) Chi-Squared Tests

The feature selection was performed on a 99-1 % split and the training set was used for the Sequential Feature Selector (this algorithm looks only at the features X, not the desired outputs y). Both selectors used a Cross-validation (CV) of 5.

A. Forward Selection

The forward selection algorithm selected 8 features in a greedy fashion [TABLE II].

TABLE II
Forward Selection of Features

Variable	Description
RIAGENDR	Gender
DMDEDUC2	Education level Adults 20+
INDFMIN2	Annual family income
FIAPROXY	Proxy used during family Interview
FSD032A	Household worried running out of food
HUQ090	Seen mental health professional/ past yr
SLQ050	Ever told doctor had trouble sleeping?
SLQ120	How often feel overly sleepy during day?

B. Backward Elimination

Backward elimination selected 8 features with only 1 being different than the forward selected features (shown in bold).

C. Chi-Squared Tests

The Chi-Squared Test tests for independence between the features and the response. The higher the ch-squared statistic, the higher the dependence between the feature and the response. The table below shows the selected features using the chi-squared test.

TABLE III
BACKWARD ELIMINATION OF FEATURES

Variable	Description
RIAGENDR	Gender
DMDEDUC2	Education level Adults 20+
RIDRETH1	**Ethnicity**
FIAPROXY	Proxy used during family Interview
FSD032A	Household worried running out of food
HUQ090	Seen mental health professional/ past yr
SLQ050	Ever told doctor had trouble sleeping?
SLQ120	How often feel overly sleepy during day?

TABLE IV
CHI-SQUARED SELECTION OF FEATURES

Variable	Description
RIAGENDR	Gender
RIDAGEYR	Age
SMQ040	Do you now smoke cigarettes?
OCQ180	Hours worked last week at all jobs
PAD680	Minutes sedentary activity
BMXBMI	Body Mass Index (kg/m**2)
DMDCITZN	Citizenship status
DMDEDUC2	Education level - Adults 20+
ALQ130	Avg # alcohol drinks/day (past 12 months)

IV. MODEL SELECTION

Given that our outcome variable may be represented numerically or categorically, we trained 2 different types of models; a Linear regression model to model the numerical outcome, and a Logistic regression model to model the categorical variable.

To remedy the imbalance in our data, we considered Synthetic Minority Oversampling Technique (SMOTE). However, we chose not to apply SMOTE to our data as it can decrease model performance for high-dimensional data, as well as decrease the separation between classes thereby introducing noise [5]. Instead, we chose to apply a weighted Logistic Regression model which is useful in modeling rare events. The main drawback of this approach was the trade-off between accuracy and recall, but we chose to prioritize recall in order to minimize the misclassification rate of respondents with depressive symptoms.

A. Linear Regression

Our initial attempt to model the relationship using a linear model, resulted in some fundamental shortcomings of linear modeling. For one, our sample was severely skewed, with only around 30 observations with an depression severity score of 20+. The rest of the 5000+ observations have a majority with 0 as their indicator value. The results of this imbalance is a very small r-squared value for all linear models.

Three linear regression models were created, all on 70-30 % split data. Each model used the selected feature sets from the earlier feature selection.

Our results showed that for linear regression the highest r-squared and lowest MSE is the linear model with forward selected features. However, with an r-squared of 0.24, only 24% of the variance can be explained by linear regression.

TABLE V
LINEAR REGRESSION PERFORMANCE

Model	MSE	RMSE	r-Squared
All Features	15.15	3.89	0.15
Forward Selection	13.09	3.61	0.26
Backward Selection	13.52	3.67	0.24

B. Logistic Regression

Two Logistic Regressions were fit; a simple unweighted model and a weighted model. Optimal weights were found using a grid search based on recall performance. The results are shown below.

TABLE VI
LOGISTIC REGRESSION PERFORMANCE

Model	AUCROC	Accuracy	Recall
Unweighted Binary LR	0.50	0.84	0.002
Weighted Binary LR	0.52	0.33	0.87

The weighted model had the best recall of 0.87, but this came at a loss for accuracy (0.33 compared to 0.84 in the unweighted model). Due to the nature of our project, we decided to prioritize recall and make our conclusions based on the weighted Logistic Regression model. The model coefficients and p-values are shown in the figure below.

	coef	std err	z	P>|z|	[0.025	0.975]
SMQ040	-7.417e-05	0.000	-0.424	0.671	-0.000	0.000
OCQ180	-4.573e-05	0.000	-0.388	0.698	-0.000	0.000
PAD680	0.0002	8.09e-05	1.967	0.049	5.74e-07	0.000
BMXBMI	-0.0251	0.010	-2.431	0.015	-0.045	-0.005
INDHHIN2	0.0022	0.004	0.590	0.555	-0.005	0.010
RIAGENDR	0.1425	0.222	0.642	0.521	-0.293	0.578
RIDAGEYR	-0.0212	0.006	-3.725	0.000	-0.032	-0.010
DMDCITZN	-0.3210	0.304	-1.055	0.292	-0.917	0.276
DMDEDUC2	-0.0089	0.005	-1.767	0.077	-0.019	0.001
ALQ130	0.0003	0.000	0.752	0.452	-0.000	0.001

Fig. 5. Logistic Regression Model Summary

V. DIRECTED ACYCLIC GRAPHS

An important part of our project, was to use DAG's to investigate the relationship between exposure variables, and outcome variables. In this case, a DAG was created showing possible causal relationships that exist between our target variable (DPQ Score), and the selected features. The green circle indicates an exposure variable. The white circles show pathways that have been conditioned on to not bias our primary outcome variable.

A. DAG #1

Education level - Depression Severity

With education level as an exposure variable, Annual family income needs to be conditioned on, to eliminate biasing pathways.

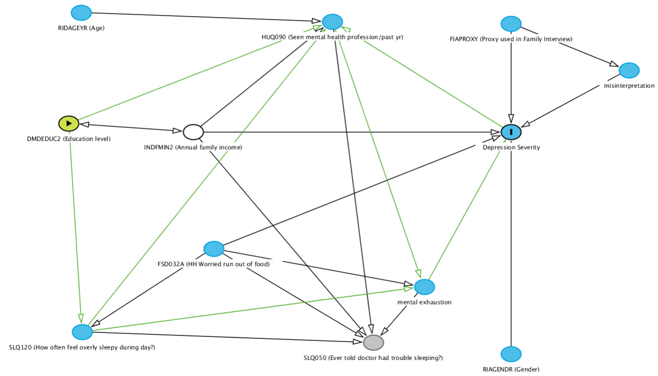
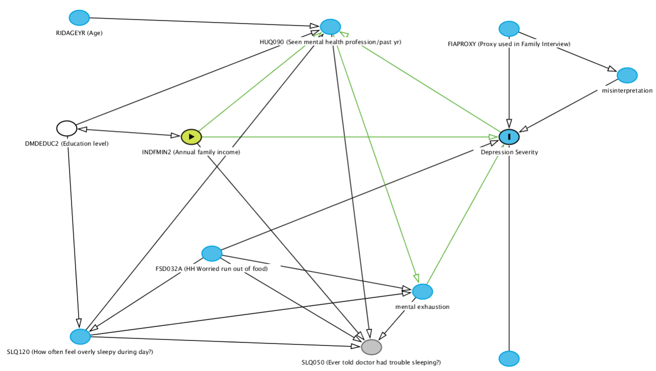


Fig. 6. DAG 1

B. DAG #2

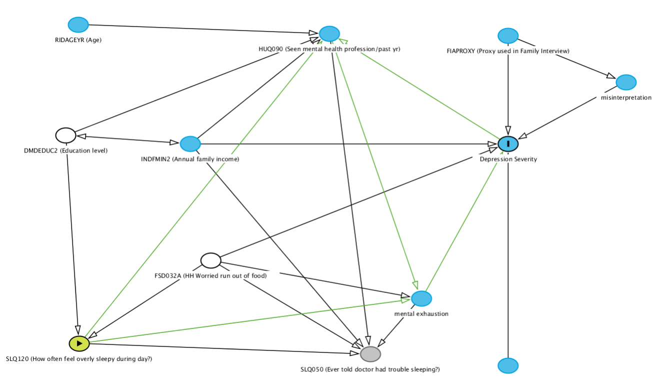
Family Income - Depression Severity



Using family income as an exposure variable, education level became a confounding variable that needs to be conditioned on to eliminate the biasing of Depression severity.

C. DAG #3

SLQ120 (How often feel sleepy during day) - Depression Severity



This variable as an exposure pathway, introduces Education level and Household food security as confounding variables that need to be accounted for.

Conclusion

Our analysis of NHANES data resulted in multiple key covariates that show importance in the detection of severe depression in survey participants. In particular, these were Age, BMI, Physical Activity level, and Education Level. Due to the shortcomings of our model, we were unable to make conclusions about the direction of causality and weight of individual covariates. However, simply identifying important risk factors for depression is a significant starting point towards further research.

Using our resulting analysis, we identify physical activity as an important variable in someone's life. It is unclear whether exercise is a causal factor, however, its relationship to depression severity cannot be ignored.

There are a few shortcomings of our data and models to keep in mind. Firstly, the NHANES Survey methodology utilizes oversampling techniques to account for underrepresented population groups in the US, and so the use of corresponding sampling weights is suggested for any NHANES data analysis. Due to time constraints, we were unable to factor this into our model design. Secondly, there were some correlated features represented in our model which ideally would be represented using interaction terms. In the future we hope to account for these shortcomings as well as explore other methods of data imputation and a combination of undersampling and oversampling techniques to model imbalanced data.

The results communicated here are a reference point for tackling the difficult epidemic of depression that continues to affect a significant proportion of American adults, and may help inform intervention strategies.

References

[1] Paul Allison, "Logistic Regression for Rare Events", https://statisticalhorizons.com/logistic-regression-for-rare-events/
[2] Zhao, L et al, "Comparison of logistic regression and linear regression in modeling percentage data." Applied and environmental microbiology vol. 67,5 (2001): 2129-35. doi:10.1128/AEM.67.5.2129-2135.2001
[3] Brody et al, "Prevalence of Depression Among Adults Aged 20 and Over: United States, 2013–2016", https://www.cdc.gov/nchs/products/databriefs/db303.htm/
[4] NHANES, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017
[5] https://www.fromthegenesis.com/smote-synthetic-minority-oversampling-technique/