Exploring Diversity in Depression within the United States

STAT 112 | Prof. Esfandiari
Department of Mathematics and Statistics, UCLA

Juliet Aguh (105913276), Joy Chen (806000839), Eric Du (905904904), Rory Freck (306145948), George Longridge (506513652), Defne Tanyildiz (105984216)

December 9th, 2024

Table of Contents

[1] Abstract

According to the National Institute of Mental Health, roughly 21 million U.S. adults experience at least one major depressive episode annually. Depression is a serious disease that can amplify the devastating effects of physical illnesses (Gaynes et al., 2002). This highlights the critical need to identify factors contributing to depression and ways to address it. To understand underlying factors of depression and other illnesses alike in the United States, the CDC and National Center for Health Statistics (NCHS) conduct annual surveys to provide up-to-date data on the prevalence, impact, and treatment of illnesses and disabilities in the U.S. These surveys are critical for informing public health strategies and policies to address mental health challenges.

Our team's methodology for analyzing predictors of depression is R code, via Random Forest, Logistic Regression, analysis of ANOVA Table, and Chi-Squared. Additionally, to ensure that our dataset met all Model Assumptions, we also leveraged Leverage Plots, Pearson Goodness of Fit test, and Contingency tables to demonstrate no multicollinearity, linear relationships between variables, and independence of errors.

The current literature has a strong focus on genetic predispositions and childhood trauma as risk factors for depression (Bembnowska & Jośko-Ochojska, 2015, pp. 117-118). However, this study looks at how factors like demographics, socioeconomic status, health, and experiences of discrimination contribute to depression risk using data from the National Health Interview Survey (NHIS). We started with eight predictors in a logistic regression model and refined it to seven for clearer interpretation. We found that financial insecurity, discrimination, and poor mental health history are key drivers of depression. To address these issues, we recommend policies that improve financial stability, reduce systemic discrimination, and expand access to mental health care. Future research should explore how overlapping factors, like disability and poverty, affect mental health and test the effectiveness of targeted interventions.

[2] Dataset Summary

Dataset Evaluated: nhis-23.csv
Codebook: nhis-adult23-codebook.pdf (2023 National Health Interview Survey (NHIS))

The NHIS 2023 Adults dataset provides national survey responses on adult health, socio-demographic characteristics, mental health, and experiences of discrimination. It offers insights into U.S. adults' health status, life satisfaction, mental health conditions, and social determinants of health. Key variables analyzed include predictors such as race, education, mental health history, and socio-economic factors, with a focus on their relationship to depression and anxiety screening outcomes. The dataset includes 28260 observations, where 1988 samples are at risk for depression (1), and 26635 samples are not at risk for depression (0).

[3] Research Questions

1. What demographic factors (e.g., race, education level) are most strongly associated with the prevalence of major depressive episodes in U.S. adults, as measured by the PHQ-2 screener?
2. How do feelings of discrimination influence the risk of depression in different demographic groups?

[4] Objective

To understand how discrimination experiences, socio-economic factors, demographic characteristics, and past mental health diagnoses influence depressive symptoms in adults, as measured by the PHQ-2 (a two-question screening tool used to identify depression).

[5] Variables Explored **

In this paper we will be exploring the effects of demographic features (race, sex), socioeconomic factors (education level, income to poverty ratio), health history (disability status, anxiety risk (yes/no)), and experiences of discrimination (feeling disrespected, feeling threatened) for an adult in the United States to be at risk of depression (yes/no).

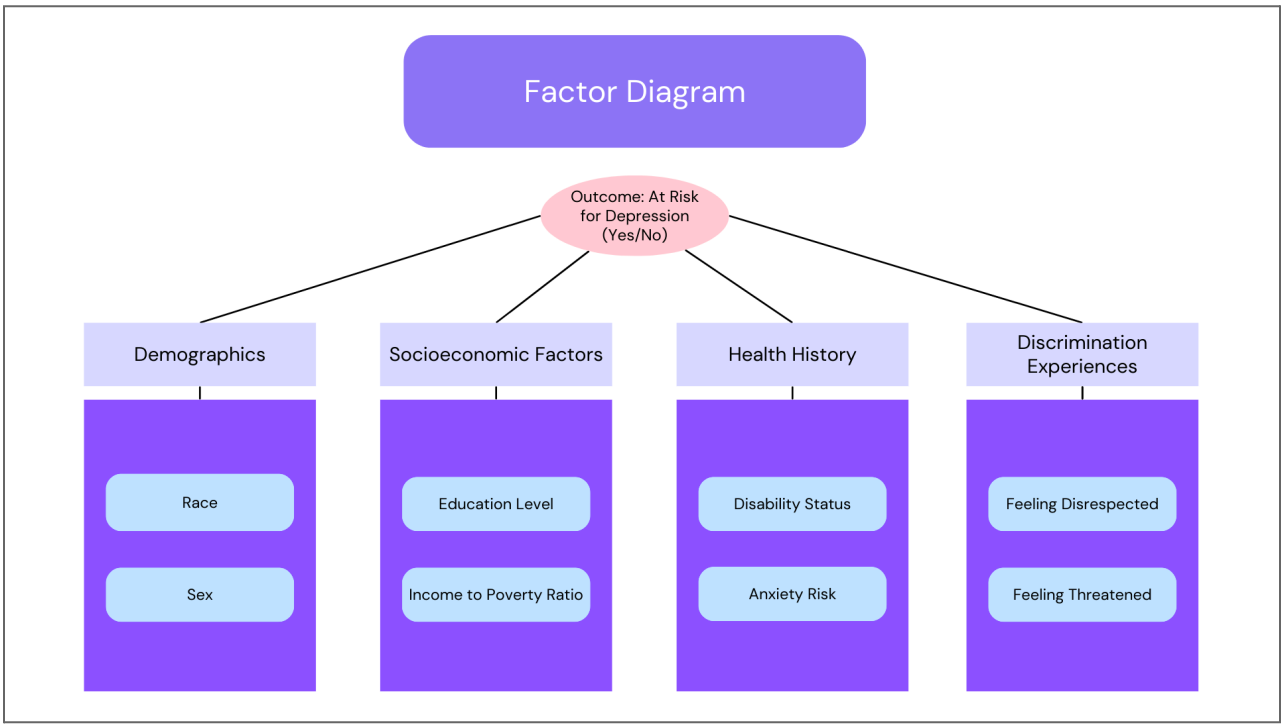

Figure 1. Variables Schematic

** *See Appendix A for Variables Codebook*

[6] Exploratory Data Analysis

To first understand and analyze the dataset, we conducted Exploratory Data Analysis using a variety of visualizations. These plots allowed us to assess the overall structure of the data and identify any imbalances within the sample. The following bar charts demonstrate the counts for the corresponding variables of this dataset used (see Figure 1.1 and 1.2) and the frequency of the predictors in relation to the response (see Figures 2.1 and 2.2).

Our analysis revealed that the dataset is heavily skewed toward certain demographic groups. The majority of individuals in the sample are white, non-disabled, treated with respect, not threatened, not impoverished, non-college-educated, and have no clinical risk of depression or anxiety. Among all variables analyzed, only the distribution of sex was balanced across the sample.

This skewed distribution poses challenges for the modeling process, as groups with smaller sample sizes are at a higher risk of being misclassified. This issue is particularly concerning for the response variable, depression risk, where the imbalance could lead to reduced sensitivity in detecting individuals at risk. Consequently, we anticipate that our model may perform less effectively in identifying cases of depression risk compared to non-risk cases, which we will address in subsequent analyses.

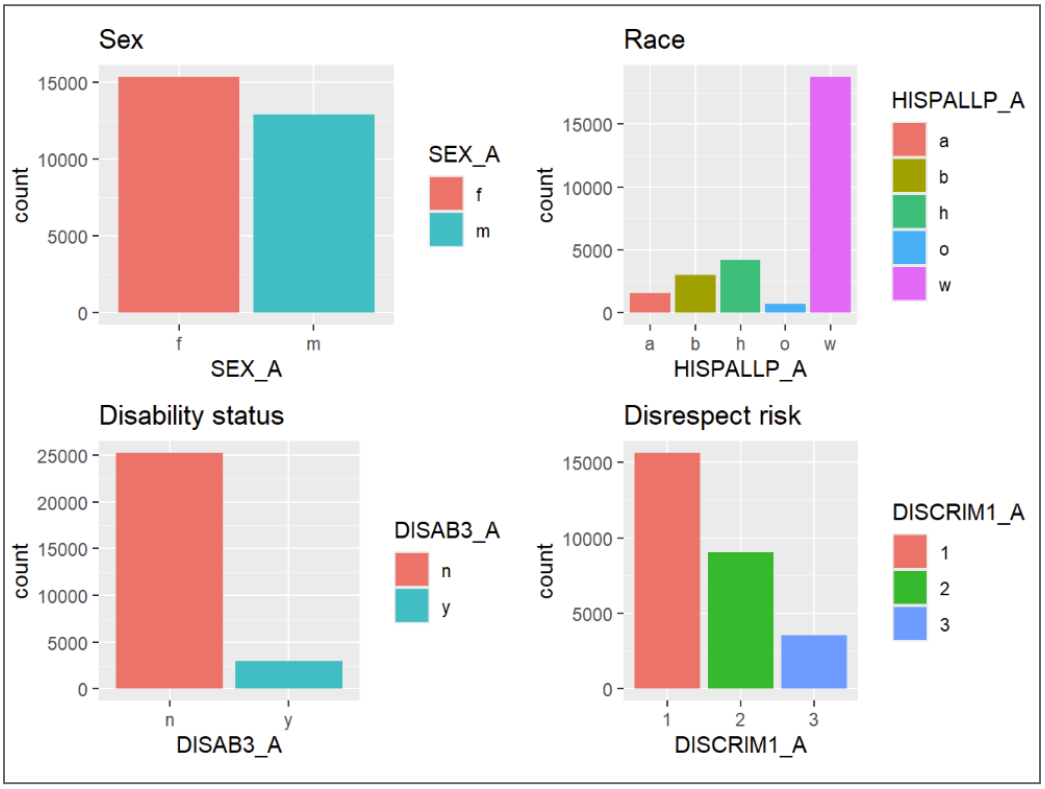

Figure 1.1. Bar Charts for Counts

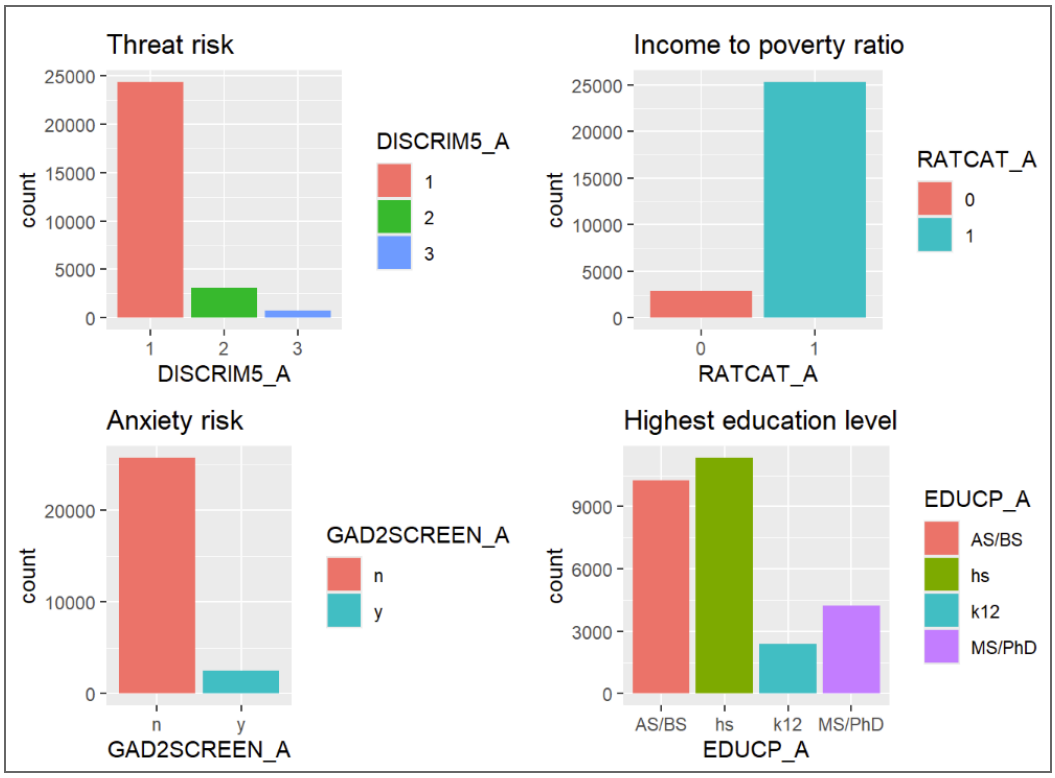

Figure 1.2. Bar Charts for Counts

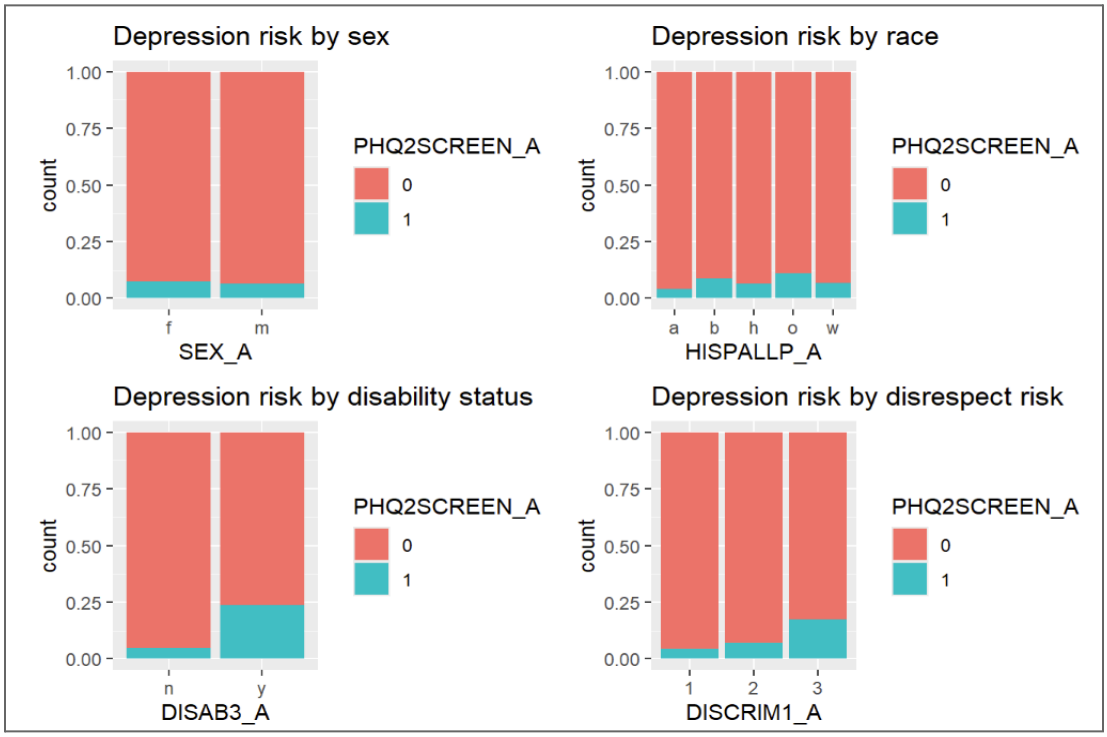

Figure 2.1. Split Bar Charts for Frequency

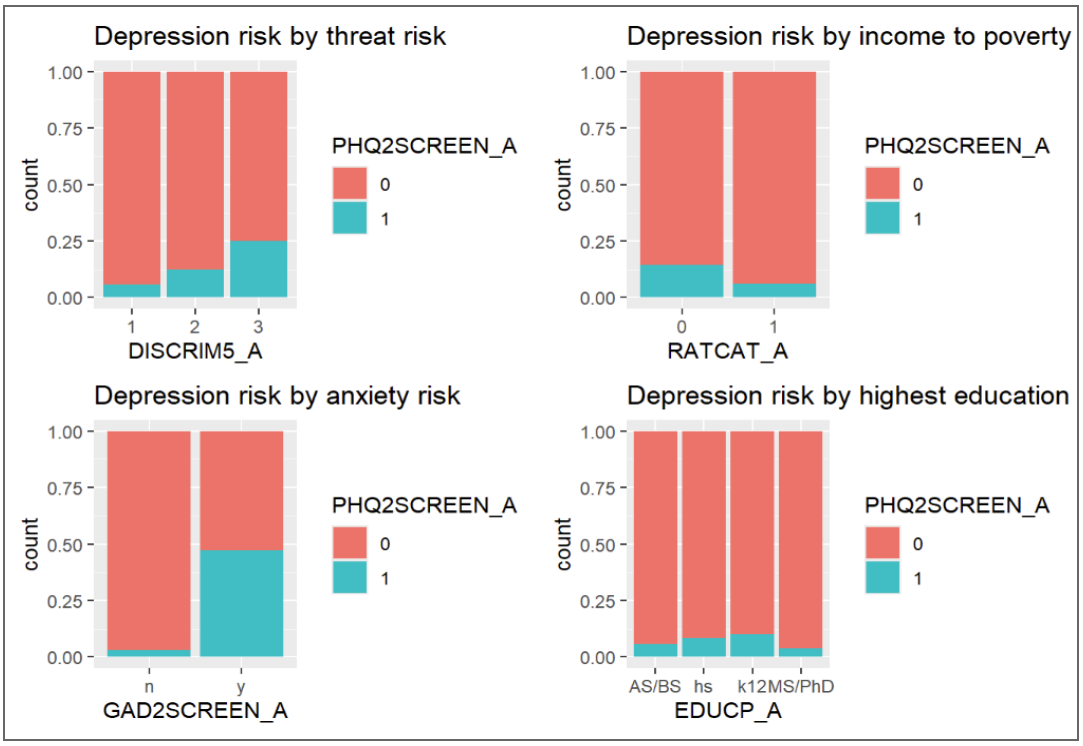

Figure 2.2. Split Bar Charts for Frequency

Additionally, we were also curious to see the distribution of samples that indicated risk of depression (yes/1) and did not indicate risk of depression (no/0). This analysis provided insight into the imbalance within the response variable itself. As expected, the majority of samples fell into the "no risk" category, further highlighting the inherent skewness of the dataset. (See Figure 3) This disproportionate distribution highlights the potential challenges our model may face in accurately identifying individuals with a risk of depression, as the minority class (yes/1) is likely to be underrepresented. Addressing this imbalance is critical for improving the sensitivity and overall performance of our predictive model, especially because the objective of our study is to understand how factors influence risk of depression.

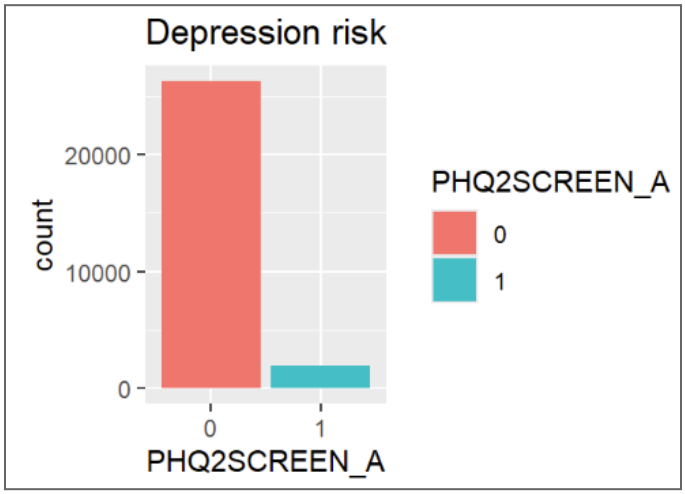

Figure 2.3 Bar Chart for Depression

[7] Data Preprocessing

Before training our models, we employed many necessary data preprocessing steps.

The original dataset was in a large CSV file with hundreds of variables. To narrow our focus, we carefully selected the variables most relevant to our research objectives. Each chosen variable was created into a factor. The levels of most of these factors were chosen by us in order to simplify the readability of the model coefficients and maximize the counts for important levels.

Furthermore, a critical step in this process involved addressing missing values (NAs), which were omitted to ensure the accuracy and reliability of our analysis. Retaining NAs could have introduced bias or inaccuracies in our results, as many analytical methods assume complete data for optimal performance. By removing these incomplete entries, we minimized potential distortions in our findings and enhanced the consistency of the dataset. After this data cleaning process, we retained 28,260 observations from the original 29,522 observations.

[8] Random Forest

We first employed Random Forest to help identify significant predictors through a non-parametric approach. This method highlighted anxiety, disability, and discrimination as the top contributors to depression risk, consistent with findings from logistic regression. Anxiety (GAD2SCREEN_A) emerged as the most significant predictor, emphasizing the co-occurrence of anxiety and depression due to shared stress-response mechanisms. Disability status (DISAB3_A) also strongly predicted depression, reflecting the compounded challenges of stigma, limited mobility, and healthcare barriers faced by disabled individuals.

Discrimination variables, such as DISCRIM1_A (being treated with less courtesy or respect) and DISCRIM5_A (feeling threatened or harassed), ranked fairly highly in variable importance. These findings reinforce the role of chronic psychosocial stressors in exacerbating depressive symptoms. Additionally, the income-to-poverty ratio (RATCAT_A) illustrated the economic dimension of depression, with financial instability contributing to heightened vulnerability. The random forest model's ability to capture nonlinear relationships and interactions, such as the combined effects of low income and discrimination, adds depth to the analysis and strengthens the confidence in these predictors.

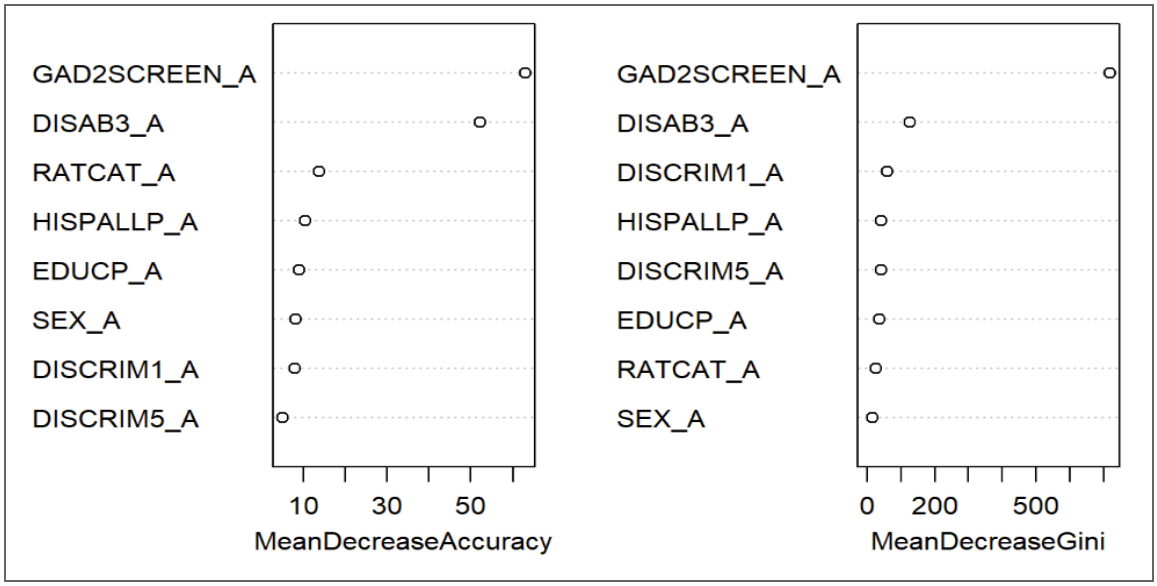

Figure 3. Random Forest Model

In summary, the Random Forest Model found that the most significant predictors of risk of depression are GAD2SCREEN_A (anxiety screening), DISAB3_A (disability indicator), RATCAT_A (family income-to-poverty ratio), DISCRIM1_A (being treated with less courtesy or respect), and HISPALLP_A (race classification). (See Figure 3)

[9] Logistic Regression

In addition to the insights gained from the Random Forest model, we proceeded to further analyze the data using Logistic Regression. Unlike Random Forest, which is a non-parametric ensemble method designed to handle complex relationships and interactions between variables, Logistic Regression is a parametric model that assumes a linear relationship between the predictors and the log-odds of the outcome. This difference in methodology allows Logistic Regression to provide clear estimates of statistical significance and effect sizes for each variable.

Interestingly, our Logistic Regression analysis revealed a key divergence from the Random Forest results. In comparison to the Random Forest Forest, our Logistic Regression Model indicated the variable HISPALLP_A (Hispanic origin and race classification of the sample adult) was not statistically significant. Thus, the two models we will be using for this study are Model 1 (full model) and Model 2 (full model excluding the HISPALLP_A variable).

In the first logistic regression model, which was the full model, HISPALLP_A was included to examine potential disparities in depression risk across racial and ethnic groups. However, its high p-value indicated that it was not statistically significant in predicting depression when controlling for other variables. This suggests that race and ethnicity, while important social determinants of health, may not independently contribute to the odds of

depression in this dataset once factors such as socioeconomic status (RATCAT_A) and discrimination (DISCRIM1_A, DISCRIM5_A) are accounted for. (See Figure 4)

```
Call:
glm(formula = PHQ2SCREEN_A ~ ., family = binomial(), data = nhis1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.956273   0.171617 -23.053  < 2e-16 ***
RATCAT_A1     -0.327604   0.078614  -4.167 3.08e-05 ***
HISPALLP_Ab    0.298565   0.164212   1.818 0.069038 .
HISPALLP_Ah    0.064831   0.162626   0.399 0.690152
HISPALLP_Ao    0.353662   0.209263   1.690 0.091020 .
HISPALLP_Aw    0.002811   0.148182   0.019 0.984865
SEX_Am         0.187701   0.057303   3.276 0.001054 **
DISAB3_Ay      1.295521   0.065977  19.636  < 2e-16 ***
DISCRIM1_A2    0.299651   0.068395   4.381 1.18e-05 ***
DISCRIM1_A3    0.690170   0.079094   8.726  < 2e-16 ***
DISCRIM5_A2    0.322855   0.079880   4.042 5.31e-05 ***
DISCRIM5_A3    0.507939   0.120734   4.207 2.59e-05 ***
EDUCP_Ahs      0.321210   0.065391   4.912 9.01e-07 ***
EDUCP_Ak12     0.392018   0.103495   3.788 0.000152 ***
EDUCP_AMS/PhD -0.187820   0.102770  -1.828 0.067614 .
GAD2SCREEN_Ay  3.009378   0.058848  51.138  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14158.1  on 28259  degrees of freedom
Residual deviance:  9673.6  on 28244  degrees of freedom
AIC: 9705.6

Number of Fisher Scoring iterations: 6
```

Figure 4. Model 1 Summary

The insignificance of the HISPALLP_A variable could also be attributed to multicollinearity, as race and ethnicity often overlap with other predictors like income, education, and experiences of discrimination. For instance, minority groups may disproportionately experience lower income levels or higher rates of harassment, which are already captured in other variables within the model. As a result, the unique contribution of HISPALLP_A became redundant or obscured in the full model.

To address this issue and enhance the interpretability of the results, a second logistic model was created, excluding the HISPALLP_A variable. This adjustment allowed for a clearer understanding of the remaining predictors' impact on depression without potential confounding from multicollinearity. The simplified model retained its robustness, with key variables like anxiety (GAD2SCREEN_A), disability (DISAB3_A), and income-to-poverty ratio (RATCAT_A) continuing to show strong associations with depression. (See Figure 5) By excluding an insignificant variable, the second model not only refined the analysis but also reinforced the importance of focusing on predictors with the most direct influence on depressive symptoms.

```
Call:
glm(formula = PHQ2SCREEN_A ~ SEX_A + DISAB3_A + DISCRIM1_A +
    DISCRIM5_A + RATCAT_A + GAD2SCREEN_A + EDUCP_A, family = binomial(),
    data = nhis1)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.88594    0.10266 -37.854  < 2e-16 ***
SEX_Am         0.18558    0.05725   3.241  0.00119 **
DISAB3_Ay      1.28780    0.06562  19.624  < 2e-16 ***
DISCRIM1_A2    0.31294    0.06824   4.586 4.52e-06 ***
DISCRIM1_A3    0.71123    0.07884   9.021  < 2e-16 ***
DISCRIM5_A2    0.32953    0.07980   4.129 3.64e-05 ***
DISCRIM5_A3    0.51953    0.12060   4.308 1.65e-05 ***
RATCAT_A1     -0.35559    0.07817  -4.549 5.39e-06 ***
GAD2SCREEN_Ay  2.99638    0.05849  51.227  < 2e-16 ***
EDUCP_Ahs      0.33438    0.06513   5.134 2.83e-07 ***
EDUCP_Ak12     0.42166    0.10192   4.137 3.52e-05 ***
EDUCP_AMS/PhD -0.19446    0.10245  -1.898  0.05767 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14158.1  on 28259  degrees of freedom
Residual deviance:  9688.7  on 28248  degrees of freedom
AIC: 9712.7

Number of Fisher Scoring iterations: 6
```

Figure 5. Model 2 Summary

9.1) Model 1: Full Model Analysis

Model 1, which includes all eight predictors, highlights several key factors influencing depression. Anxiety (GAD2SCREEN_A) is the strongest predictor, with individuals reporting anxiety having 20.27 times higher odds of depression. This finding highlights the well-documented co-occurrence of anxiety and depression. Disability status (DISAB3_A) also plays a significant role, with disabled individuals facing 3.65 times higher odds of depression. Socioeconomic factors, such as income-to-poverty ratio (RATCAT_A), are important, as individuals above the poverty line have 0.72 times lower odds of depression compared to those with higher incomes. Discrimination (DISCRIM1_A Level 3) increases the odds of depression by 2.0 times, and individuals reporting harassment (DISCRIM5_A Level 3) have 1.7 times higher odds of depression. The HISPALLP_A (race classification) variable did not show a statistically significant relationship with depression. The confidence intervals for the odds ratios for all racial categories (e.g., Black: OR 1.35, 95% CI: [0.98, 1.85], p = 0.069; Hispanic: OR 1.07, 95% CI: [0.79, 1.45], p = 0.690) included 1, suggesting no meaningful difference in depression risk by race.

9.2) Model 2: Simplified Model Analysis

Model 2, excluding HISPALLP_A, retained anxiety, disability, and socioeconomic factors as the strongest predictors of depression. Anxiety (GAD2SCREEN_A) continues to be the most significant predictor, with individuals reporting anxiety having 20.0 times higher odds of depression. Disability status (DISAB3_A) remains significant, with disabled individuals facing 3.62 times higher odds of depression. Individuals above the poverty line have 0.70 lower odds of depression. Discrimination also continues to significantly affect depression risk, with those reporting discourteous treatment (DISCRIM1 Level 3) having 2.04 times higher odds of depression and harassment (DISCRIM5 Level 3) having 1.68 times higher odds of depression. In terms of education, those with 12 years of schooling have 1.52 times higher odds of depression, while those with advanced degrees (MS/PhD) have lower odds of depression, but this result was not statistically significant. The exclusion of HISPALLP_A led to weaker associations for some predictors, indicating that race may have acted as a confounder or moderator in Model 1. This change suggests that excluding race oversimplifies the complex relationships between race, socioeconomic status, and discrimination, which could have influenced the depression risk.

9.3) Accuracy of Models: A Comparison

The accuracy of the two logistic regression models was evaluated using pseudo R-squared values and the area under the curve (AUC) from the ROC analysis. These metrics offer insights into the models' ability to predict depression and differentiate between individuals with and without depressive symptoms.

For Model 1, the pseudo R-squared value was **0.3167429**, indicating that approximately 31.7% of the variation in depression risk was explained by the predictors. While pseudo R-squared values tend to be lower than traditional R-squared values in linear regression, this result suggests a strong explanatory power for a logistic model. The AUC for Model 1 was **0.8623**, signifying excellent performance in distinguishing between depressed and non-depressed individuals. With an AUC value above 0.85, the model demonstrated a high degree of accuracy, minimizing the likelihood of misclassification. (See Figure 6)

The confusion matrix for Model 1 further supports this, as the model correctly classified 25,962 individuals as not depressed and 634 individuals as depressed. However, it did misclassify 1,310 individuals as depressed when they were not and 354 individuals as not depressed when they were. The total error rate was 0.0589. The specificity (true positive) rate was 0.9865 and the sensitivity (true negative) rate was 0.3261. (See Figure 7)

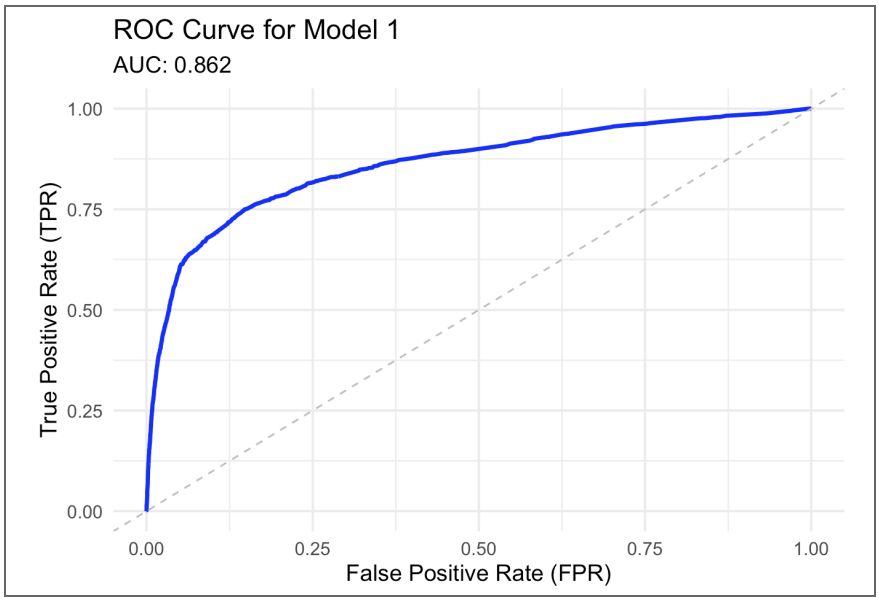

Figure 6. Model 1 Area Under Curve

Model 1	Observed 0	Observed 1
Predicted 0	25962	1310
Predicted 1	354	634

Figure 7. Model 1 Confusion Matrix

Model 2, which excluded the HISPALLP_A variable due to its high p-value and insignificance, showed similar performance. The pseudo R-squared for Model 2 was **0.3156796**, slightly lower than Model 1 by approximately 0.001. This minimal difference suggests that removing HISPALLP_A had little effect on the model's ability to explain variance in depression risk. The AUC for Model 2 was **0.8618**, nearly identical to Model 1, further demonstrating that the exclusion of this variable did not significantly impact the model's predictive power. (See Figure 8) In terms of classification, Model 2 correctly identified 25,969 individuals as not depressed and 614 individuals as depressed. It misclassified 1,330 individuals as depressed when they were not and 347 individuals as not depressed when they were. The total error rate was 0.0593. These results are nearly equivalent to those of Model 1, with slight variations in the number of false positives and false negatives.

The specificity (true positive) rate was 0.9868 and the sensitivity (true negative) rate was 0.3158. (See Figure 9)

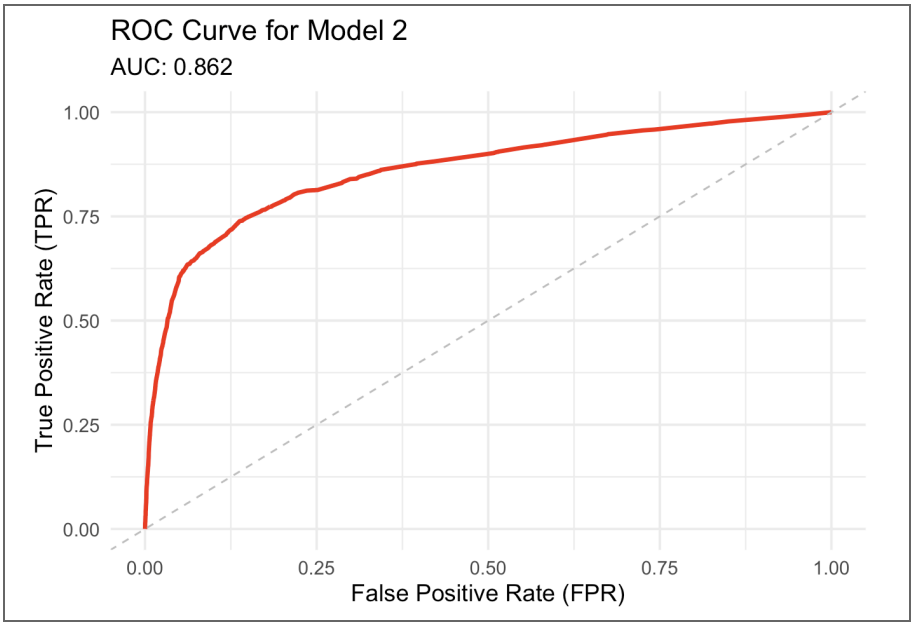

Figure 8. Model 2 Area Under Curve

Model 2	Observed 0	Observed 1
Predicted 0	25969	1330
Predicted 1	347	614

Figure 9. Model 2 Confusion Matrix

When comparing the two models, both perform exceptionally well, with nearly identical pseudo R-squared values and AUC scores. While Model 1 provided marginally better fit, as evidenced by its slightly higher pseudo R-squared and AUC, the differences are minimal. This indicates that HISPALLP_A did not meaningfully contribute to the model's predictive performance. From a practical standpoint, Model 2 offers the advantage of simplicity and ease of interpretation by excluding an insignificant variable. This streamlined approach is particularly valuable in applied settings where efficiency is prioritized without sacrificing accuracy.

[10] Plot of Odds

To further validate our reasoning for removing the HISPALLP_A variable in our reduced model, we plotted the odds of both models for comparison.

When plotting the odds for Model 1, we observed that the confidence intervals for the HISPALLP_A variable all included 1. (See Figure 10) This indicates that the odds of the outcome were not significantly different across the levels of race classification. In other words, race classification did not appear to substantially influence the likelihood of the outcome, supporting its exclusion from the reduced model.

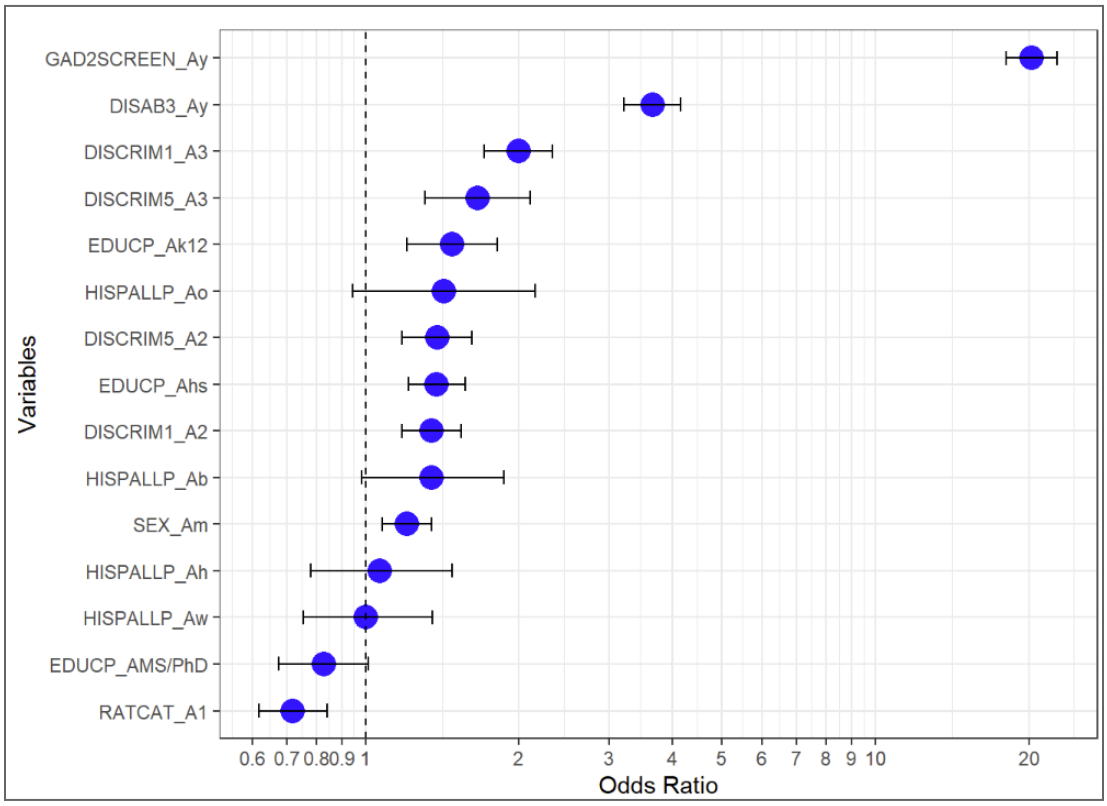

Figure 10. Model 1 Plot of Odds

To address this further, we examined the odds plot for Model 2, the reduced model. (See Figure 11) In comparison, none of the variables in Model 2 had confidence intervals that included 1. This suggests that the removal of HISPALLP_A allowed for a more refined model where all remaining predictors showed a stronger association with the outcome, enhancing the clarity of their contributions. This result provides additional justification for excluding HISPALLP_A from the reduced model while ensuring the remaining predictors retained their significance.

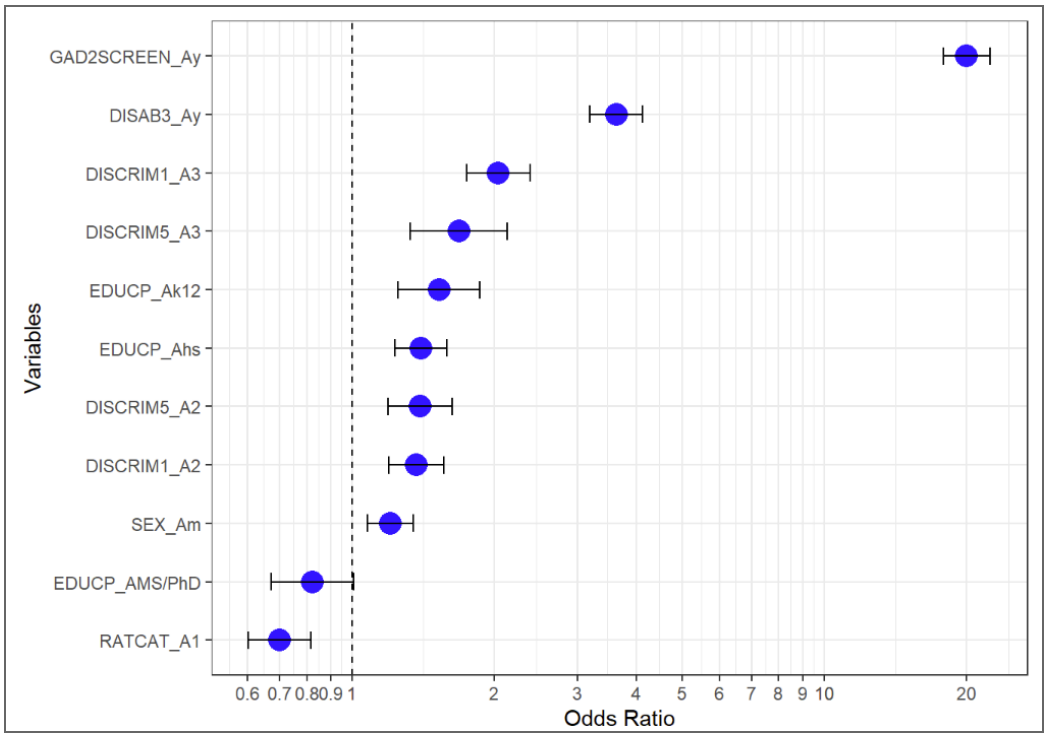

Figure 11. Model 2 Plot of Odds

By comparing the odds plots of Model 1 (the full model) and Model 2 (the reduced model), we found that removing HISPALLP_A had minimal impact on the estimated odds for other predictors. This reinforces our conclusion that the variable does not independently contribute significant predictive value. However, the observed limitations in performance metrics such as AIC and ANOVA suggest that HISPALLP_A may interact with other predictors in ways that are not captured in our reduced model.

These findings underscore the importance of considering interaction effects in future analyses. Incorporating interaction terms could provide a more comprehensive understanding of the relationships between predictors and ensure that critical information about variables like race classification is not inadvertently lost.

[11] Model Assumptions

For the Logistic Regression Model, we also had to consider various model assumptions as follows.

11.1) Large sample size:

First, we needed to account for the assumption of a large sample size to ensure that the model's estimated coefficients are statistically reliable. Small sample sizes can lead to overfitting or underfitting, making the model less generalizable to new data.

This assumption was met due to our dataset containing 28,260 observations.

11.2) Linear relationship between between the predictors and the response:

Another critical assumption is that there is a linear relationship between the predictors and the log-odds of the outcome variable. This means that the effect of each predictor on the response variable is assumed to be additive and constant, which may not always reflect the true complexity of the data.

This assumption was met due to significant p-values in logistic regression and Pearson's Test for Goodness of Fit.

11.2.a.) Pearson Goodness of Fit Test:

H_0: The logistic regression model is a good fit for the data
H_a: The logistic regression model is not a good fit for the data

Chi-squared statistic: 26413.16 (sum of squared Pearson residuals)

95% quartile on 28248 degrees of freedom: 28640.1

Since the statistic is lower than the quartile value, we fail to reject the null hypothesis. Model 2 fits the data well.

11.3) Multicollinearity:

Additionally, another critical assumption is that there is no significant multicollinearity among the predictors. If predictors are highly correlated, it becomes difficult to distinguish their individual effects on the outcome, potentially leading to unstable coefficient estimates.

This assumption was unmet due to significant relationships between variables in chi-squared goodness of fit tests. While variance inflation factor (VIF) is typically used to assess multicollinearity in linear regression, we hesitated to use it in logistic regression due to its dependence on r-squared. Logistic regression can only obtain a pseudo r-squared which cannot be interpreted as r-squared (OARC Stats). Thus, we obtained VIFs, all of which were below 1.5, but we took them with a grain of salt.

11.4) Leverage and Influential Points

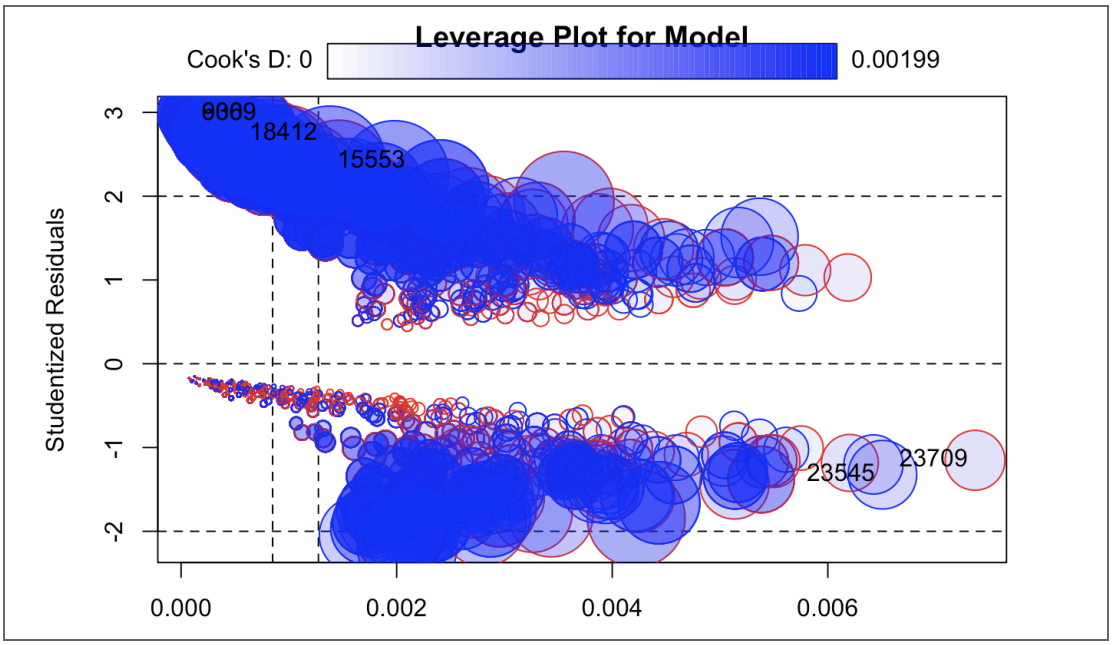

Figure 12: Leverage Plot for Model 2

The logistic model assumes that there are no influential points. The assumption is met because while there are influential points, they are not overly numerous relative to the sample size.

11.5) Independence of Errors:

Finally, the Logistic Regression Model assumes that the observations are independent of each other, meaning there should be no correlation between the residuals or errors in the model. Violation of this assumption could lead to biased estimates and affect the model's predictive accuracy.

This assumption was met due to random sampling. No two observations should be related to each other.

[12] K-Fold Cross Validation

Neither model had much of a change in accuracy or kappa in 5-fold and 10-fold cross validation. While accuracy is high, this may be due to how imbalanced the dataset is. Kappa accounts for that imbalance. A kappa around 0.4 suggests moderate agreement with the model and the data.

12.1) 5-Fold Cross Validation

Model 1: Accuracy = 0.9408, Kappa = 0.4015
Model 2: Accuracy = 0.9402, Kappa = 0.3986

12.2) 10-Fold Cross Validation

Model 1: Accuracy = 0.9410, Kappa = 0.4052
Model 2: Accuracy = 0.9406, Kappa = 0.3979

[13] Discussion

After reviewing both logistic regression models, we decided to recommend the reduced model due to its heightened simplicity and similar accuracy compared to the full model. Both models showed that all the categories that our predictors belonged to: demographics, socioeconomics, health history, and discrimination are all important in predicting one's risk of depression. Looking at the specific variables, we see some particularly interesting results. Men were significantly more likely to be at risk for depression than women. Likewise, those who had income below the poverty level were significantly more likely to be at risk than their counterparts. These findings may defy common notions such as men being unemotional and money not "buying happiness". Also, as feelings of discrimination increased, so did the risk of depression. Men, the low-income classes, and discriminated-against persons are often overlooked in societal conversations about depression and may not be taken seriously about their worries over mental health. Mental health stigma may be the cause of this lack of societal acknowledgement, and stigma may lead to reduced physical health and poorer interpersonal relationships (Sickel et. al, 2014). Our research indicates that this needs to change as it is both the ethical and logical path to take.

[14] Limitations

Our study faced certain limitations, particularly concerning model assumptions, which warrant discussion. One significant limitation involved the issue of multicollinearity among predictors. Chi-squared tests performed on all combinations of predictors revealed statistically significant relationships between them, providing strong evidence of multicollinearity. This high degree of intercorrelation among predictors can distort the estimated coefficients in models like Logistic Regression, making it difficult to accurately interpret the individual effects of each variable.

Additionally, our reduced logistic regression model, which excluded the variable HISPALLP_A (race classification) due to its insignificance in the full model, underperformed compared to the full model. Metrics such as accuracy, Akaike Information Criterion (AIC), and ANOVA results indicated that the full model was superior. This suggests that while HISPALLP_A was not significant on its own, it may have interacted significantly with other predictors in the dataset. By removing this variable, the reduced model likely lost important information about these potential interactions.

This finding highlights a key limitation in our analysis: the lack of explicit interaction analysis in our modeling process. Including interaction terms might have allowed us to capture complex relationships between predictors, such as the way race could amplify or moderate the effects of variables like income-to-poverty ratio or perceived discrimination. Future studies would benefit from incorporating interaction analysis to better understand these nuanced relationships and improve the overall explanatory power of the model.

[15] Conclusion

Our analysis highlights the complex relationships between various predictors and depression, providing valuable insights through both Random Forest and Logistic Regression models. While the Random Forest model revealed intricate patterns and interactions, the Logistic Regression approach allowed for a more interpretable understanding of the statistical significance and effect sizes of key variables. Despite some divergence in findings, particularly regarding the insignificance of the HISPALLP_A variable, both models underscored the importance of anxiety, disability, socioeconomic factors, and discrimination in predicting depression.

The decision to refine our analysis with a second model excluding the HISPALLP_A variable resulted in a more streamlined and interpretable model, which reinforced the prominence of anxiety and socioeconomic stressors as major contributors to depression risk. However, we also observed the potential role of multicollinearity and the need for future models to explore interaction effects, especially related to race and ethnicity, to capture more nuanced relationships.

Overall, both models demonstrated strong predictive performance, with minor differences in accuracy and AUC, suggesting that the removal of HISPALLP_A did not significantly impact the model's explanatory power. The findings emphasize the importance of addressing mental health disparities among marginalized populations, particularly those facing discrimination and economic insecurity. Future research should consider incorporating interaction terms and refining model assumptions to enhance the depth of analysis and ensure more accurate, actionable insights in addressing depression risk.

[16] Suggestions

One of the key areas for improvement is incorporating intersectionality in future analyses. Depression does not exist in isolation but is shaped by overlapping identities and experiences, such as race, gender, socioeconomic status, and disability. Future models should include interactions between variables to better understand how these factors combine to influence depression risk. Incorporating interaction effects into the model is crucial for capturing the complexities of how various predictors influence depression risk. This approach not only enhances the model's predictive power but also provides deeper insights into intersectional vulnerabilities, enabling the development of targeted interventions and policies tailored to the needs of marginalized populations.

To move beyond correlation and explore causation, we recommend conducting longitudinal studies. These studies will allow us to track individuals over time, helping to reveal how factors like discrimination, financial insecurity, or health changes contribute to the onset or persistence of depression.

Currently, our model uses a binary classification to determine whether someone is at risk for depression. While this is useful, it simplifies a complex condition. Instead, we suggest using a continuous measure of depression severity, such as PHQ-9 scores. This approach captures the full spectrum of depression, from mild to severe, and provides a more nuanced understanding of how predictors influence varying levels of mental health.

Another critical improvement is to incorporate regional and cultural factors into the model. Depression risk and its predictors can vary significantly across geographic locations and cultural contexts. By including these factors, we can improve the generalizability and contextual relevance of our findings, making the model more applicable to diverse populations and informing localized interventions.

Our current model has a significant class imbalance, with far fewer observations for the at-risk group. To address this, we recommend assigning higher weights to the at-risk class in the logistic regression model. This ensures that the model adequately learns patterns related to this group, improving its predictive power. Additionally, we propose applying survey weights to correct for over- or under-sampled populations, ensuring our results reflect the broader population accurately.

Economic insecurity, as measured by the income-to-poverty ratio, was a critical driver of depression, highlighting the need for policies that enhance financial stability. Programs offering affordable housing, unemployment benefits, and food security can alleviate financial stress and reduce depressive symptoms.

Similarly, the strong association between discrimination and depression underscores the importance of addressing systemic inequities through workplace diversity initiatives, anti-discrimination laws, and community education campaigns.

The high prevalence of co-occurring anxiety and depression calls for integrated mental health services that address both conditions. Early screening for anxiety and depression using screeners like the two used in our data, combined with holistic treatment approaches, can improve outcomes.

Moreover, tailored programs especially for disabled individuals, such as enhanced accessibility and destigmatization efforts, are essential to mitigate their elevated depression risk.

[17] References

[1] Bembnowska, M., & Jośko-Ochojska, J. (2015). What causes depression in adults. *Polish Journal of Public Health*, *125*(2), 116-120.

[2] *FAQ: What are pseudo R-squareds?*. OARC Stats. (2011, October 20). https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/#:~:text=These%20are%20%E2%80%9Cpseudo%E2%80%9D%20R%2D,commonly%20arise%20with%20raw%20likelihood.

[2] Gaynes, B. N., Burns, B. J., Tweed, D. L., & Erickson, P. (2002). Depression and health-related quality of life. *The Journal of nervous and mental disease*, *190*(12), 799-806.

[4] "Major Depression." National Institute of Mental Health, U.S. Department of Health and Human Services, July 2023, www.nimh.nih.gov/health/statistics/major-depression

[5] Sickel, A. E., Seacat, J. D., & Nabors, N. A. (2014). Mental health stigma update: A review of consequences. *Advances in Mental Health*, *12*(3), 202-215.

Appendix

A. Variables Codebook

Predictors of Interest	Description
DISCRIM1_A	Experience of being treated with less courtesy or respect; levels indicate increasing feelings of disrespect
DISCRIM5_A	Experience of feeling threatened or harassed; levels indicate increasing feelings of being threatened
HISPALLP_A	Hispanic origin and race classification of the sample adult; Asian, White, Black, Hispanic, other
EDUCP_A	Highest educational level of the sample adult; K-12, highschool diploma, Associate's or Bachelor's degree, Master's or PhD degree
SEX_A	Sex of the sample adult; male or female
RATCAT_A	Family income-to-poverty ratio for the sample adult's family; 1 indicates at or above poverty level whilst 0 indicates below
GAD2SCREEN_A	Screener assessing the frequency of anxiety symptoms over the past two weeks; y indicates risk of anxiety whilst n indicates no risk
DISAB3_A	Disability indicator, yes or no
Outcome	**Description**
PHQ2SCREEN_A	Screener assessing frequency of depressed mood and lack of interest over the past two weeks; 1 indicates risk of depression whilst 0 indicates no risk

B. R Libraries

i) Car: used to create a leverage plot
ii) caret: used for cross validation
iii) Dplyr: used for dataset manipulation
iv) ggplot2: aids in creating EDA plots
v) gridExtra: helps arrange EDA plots
vi) OddsPlotty: generates odds ratios and plot of odds
vii) pROC: used to create ROC curves
viii) randomForest: used for random forest models and variance important plots

C. R Codes

```
nhis <- read.csv("adult23.csv")

nhis1 <- nhis[c("RATCAT_A", "HISPALLP_A", "SEX_A", "DISAB3_A", "DISCRIM1_A",
"DISCRIM5_A", "EDUCP_A", "GAD2SCREEN_A", "PHQ2SCREEN_A")]

# income to poverty ratio
nhis1$RATCAT_A <- ifelse(nhis1$RATCAT_A <= 3, 0, 1)

# race and ethnicity
nhis1$HISPALLP_A <- ifelse(nhis1$HISPALLP_A == 1, "h",
                    ifelse(nhis1$HISPALLP_A == 2, "w",
                    ifelse(nhis1$HISPALLP_A == 3, "b",
                    ifelse(nhis1$HISPALLP_A == 4, "a", "o"))))

# sex
nhis1$SEX_A <- ifelse(nhis1$SEX_A == 1, "m",
               ifelse(nhis1$SEX_A == 2, "f", NA))

# disability status
nhis1$DISAB3_A <- ifelse(nhis1$DISAB3_A == 1, "y",
                  ifelse(nhis1$DISAB3_A == 2, "n", NA))

# disrespected
nhis1$DISCRIM1_A <- ifelse(nhis1$DISCRIM1_A <= 2, 3,
                    ifelse(nhis1$DISCRIM1_A >= 3 & nhis1$DISCRIM1_A <= 4, 2,
                    ifelse(nhis1$DISCRIM1_A == 5, 1, NA)))

# threatened
nhis1$DISCRIM5_A <- ifelse(nhis1$DISCRIM5_A <= 2, 3,
                    ifelse(nhis1$DISCRIM5_A >= 3 & nhis1$DISCRIM5_A <= 4, 2,
                    ifelse(nhis1$DISCRIM5_A == 5, 1, NA)))

# education
nhis1$EDUCP_A <- ifelse(nhis1$EDUCP_A <= 2, "k12",
                 ifelse(nhis1$EDUCP_A >= 3 & nhis1$EDUCP_A <= 5, "hs",
                 ifelse(nhis1$EDUCP_A >= 6 & nhis1$EDUCP_A <= 8, "AS/BS",
                 ifelse(nhis1$EDUCP_A >= 9 & nhis1$EDUCP_A <= 10, "MS/PhD", NA))))

# anxiety disorder
nhis1$GAD2SCREEN_A <- ifelse(nhis1$GAD2SCREEN_A == 1, "y",
                      ifelse(nhis1$GAD2SCREEN_A == 2, "n", NA))
```

```
# depression risk
nhis1$PHQ2SCREEN_A <- ifelse(nhis1$PHQ2SCREEN_A == 1, "1",
                      ifelse(nhis1$PHQ2SCREEN_A == 2, "0", NA))

nhis1 <- na.omit(nhis1)

for (i in 1:ncol(nhis1)) {
 nhis1[,i] <- as.factor(nhis1[,i])
}

library(ggplot2)

p1 <- ggplot(data = nhis1, aes(x = SEX_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by sex")
p2 <- ggplot(data = nhis1, aes(x = HISPALLP_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by race")
p3 <- ggplot(data = nhis1, aes(x = DISAB3_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by disability status")
p4 <- ggplot(data = nhis1, aes(x = DISCRIM1_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by disrespect risk")
p5 <- ggplot(data = nhis1, aes(x = DISCRIM5_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by threat risk")
p6 <- ggplot(data = nhis1, aes(x = RATCAT_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by income to poverty ratio")
p7 <- ggplot(data = nhis1, aes(x = GAD2SCREEN_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by anxiety risk")
p8 <- ggplot(data = nhis1, aes(x = EDUCP_A, fill = PHQ2SCREEN_A)) + geom_bar(position = "fill") + ggtitle(label = "Depression risk by highest education level")

library(gridExtra)

grid.arrange(p1, p2, p3, p4, nrow = 2)

grid.arrange(p5, p6, p7, p8, nrow = 2)


p9 <- ggplot(data = nhis1, aes(x = SEX_A, fill = SEX_A)) + geom_bar() + ggtitle(label = "Sex")
p10 <- ggplot(data = nhis1, aes(x = HISPALLP_A, fill = HISPALLP_A)) + geom_bar() + ggtitle(label = "Race")
p11 <- ggplot(data = nhis1, aes(x = DISAB3_A, fill = DISAB3_A)) + geom_bar() + ggtitle(label = "Disability status")
p12 <- ggplot(data = nhis1, aes(x = DISCRIM1_A, fill = DISCRIM1_A)) + geom_bar() + ggtitle(label = "Disrespect risk")
p13 <- ggplot(data = nhis1, aes(x = DISCRIM5_A, fill = DISCRIM5_A)) + geom_bar() + ggtitle(label = "Threat risk")
```

```
p14 <- ggplot(data = nhis1, aes(x = RATCAT_A, fill = RATCAT_A)) + geom_bar() + ggtitle(label = "Income to poverty ratio")
p15 <- ggplot(data = nhis1, aes(x = GAD2SCREEN_A, fill = GAD2SCREEN_A)) + geom_bar() + ggtitle(label = "Anxiety risk")
p16 <- ggplot(data = nhis1, aes(x = EDUCP_A, fill = EDUCP_A)) + geom_bar() + ggtitle(label = "Highest education level")
p17 <- ggplot(data = nhis1, aes(x = PHQ2SCREEN_A, fill = PHQ2SCREEN_A)) + geom_bar() + ggtitle(label = "Depression risk")

grid.arrange(p9, p10, p11, p12, nrow = 2)

grid.arrange(p13, p14, p15, p16, nrow = 2)

grid.arrange(p17, nrow = 2, ncol = 2)


m1 <- glm(PHQ2SCREEN_A ~ ., data = nhis1, family = binomial())

summary(m1)

m2 <- glm(PHQ2SCREEN_A ~ SEX_A + DISAB3_A + DISCRIM1_A + DISCRIM5_A + RATCAT_A + GAD2SCREEN_A + EDUCP_A, data = nhis1, family = binomial())

summary(m2)


library(OddsPlotty)

odds_plot(m1)

odds_plot(m2)

library(randomForest)

rf1 <- randomForest(PHQ2SCREEN_A ~ ., data = nhis1, importance = TRUE, keep.forest = TRUE)

varImpPlot(rf1)


prob1 <- predict(m1, type = "response")

pred1 <- rep("1", nrow(nhis1))

pred1[prob1 < 0.5] <- "0"

table(predicted = pred1, actual = nhis1$PHQ2SCREEN_A)


prob2 <- predict(m2, type = "response")

pred2 <- rep("1", nrow(nhis1))

pred2[prob2 < 0.5] <- "0"
```

```
table(predicted = pred2, actual = nhis1$PHQ2SCREEN_A)


# chi-squared tests for independence
chi_all <- function(var){
  p_vals <- vector()
  names <- vector()
  for (i in 1:ncol(nhis1)) {
  empty <- vector()
  empty <- chisq.test(table(var, nhis1[,i]))$p.value
  p_vals <- c(empty, p_vals)
  }

  for (i in 1:ncol(nhis1)) {
  empty <- vector()
  empty <- colnames(nhis1)[i]
  names <- c(empty, names)
  }
  cbind(names, p_vals)
}


chi_all(nhis1$RATCAT_A)
chi_all(nhis1$HISPALLP_A)
chi_all(nhis1$SEX_A)
chi_all(nhis1$DISAB3_A)
chi_all(nhis1$DISCRIM1_A)
chi_all(nhis1$DISCRIM5_A)
chi_all(nhis1$EDUCP_A)
chi_all(nhis1$GAD2SCREEN_A)
chi_all(nhis1$PHQ2SCREEN_A)

anova(m2, m1, test = "Chisq") # goodness of fit comparing models
```

```
library(caret)

ctrl1 <- trainControl(method = "cv", number = 5)

ctrl2 <- trainControl(method = "cv", number = 10)


# model 1

train(PHQ2SCREEN_A ~ SEX_A + DISAB3_A + DISCRIM1_A + DISCRIM5_A + RATCAT_A + GAD2SCREEN_A + EDUCP_A + HISPALLP_A, data = nhis1, method = "glm", trControl = ctrl1)

train(PHQ2SCREEN_A ~ SEX_A + DISAB3_A + DISCRIM1_A + DISCRIM5_A + RATCAT_A + GAD2SCREEN_A + EDUCP_A + HISPALLP_A, data = nhis1, method = "glm", trControl = ctrl2)


# model 2

train(PHQ2SCREEN_A ~ SEX_A + DISAB3_A + DISCRIM1_A + DISCRIM5_A + RATCAT_A + GAD2SCREEN_A + EDUCP_A, data = nhis1, method = "glm", trControl = ctrl1)

train(PHQ2SCREEN_A ~ SEX_A + DISAB3_A + DISCRIM1_A + DISCRIM5_A + RATCAT_A + GAD2SCREEN_A + EDUCP_A, data = nhis1, method = "glm", trControl = ctrl2)



# Pearson's goodness-of-fit test for m1 (full model)
pearson_chisq <- sum(residuals(m1, type = "pearson")^2)
df <- m1$df.residual
qchisq(.95, df)
p_value <- pchisq(pearson_chisq, df, lower.tail = FALSE)

cat("Pearson's Chi-Squared:", pearson_chisq, "\n")
cat("Degrees of Freedom:", df, "\n")
cat("P-value:", p_value, "\n")

# Pearson's goodness-of-fit test for m2 (reduced model)
pearson_chisq2 <- sum(residuals(m2, type = "pearson")^2)
df2 <- m2$df.residual
qchisq(.95, df2)
p_value2 <- pchisq(pearson_chisq2, df2, lower.tail = FALSE)

cat("Pearson's Chi-Squared (m2):", pearson_chisq2, "\n")
cat("Degrees of Freedom (m2):", df2, "\n")
```

```
cat("P-value (m2):", p_value2, "\n")


library(pROC)
library(ggplot2)
library(dplyr)

# Generate ROC data for Model 1
roc1 <- roc(nhis1$PHQ2SCREEN_A, prob1)
roc1_df <- data.frame(
  TPR = rev(roc1$sensitivities),
  FPR = rev(1 - roc1$specificities)
)
auc1 <- auc(roc1)

# Plot ROC Curve for Model 1
roc1_plot <- ggplot(roc1_df, aes(x = FPR, y = TPR)) +
  geom_line(color = "blue", size = 1.2) +
  geom_abline(linetype = "dashed", color = "gray") +
  labs(
    title = "ROC Curve for Model 1",
    subtitle = paste("AUC:", round(auc1, 3)),
    x = "False Positive Rate (FPR)",
    y = "True Positive Rate (TPR)"
  ) +
  theme_minimal(base_size = 14)

# Generate ROC data for Model 2
roc2 <- roc(nhis1$PHQ2SCREEN_A, prob2)
roc2_df <- data.frame(
  TPR = rev(roc2$sensitivities),
  FPR = rev(1 - roc2$specificities)
)
auc2 <- auc(roc2)

# Plot ROC Curve for Model 2
roc2_plot <- ggplot(roc2_df, aes(x = FPR, y = TPR)) +
  geom_line(color = "red", size = 1.2) +
  geom_abline(linetype = "dashed", color = "gray") +
  labs(
    title = "ROC Curve for Model 2",
    subtitle = paste("AUC:", round(auc2, 3)),
    x = "False Positive Rate (FPR)",
    y = "True Positive Rate (TPR)"
  ) +
```

```
  theme_minimal(base_size = 14)

# Display both plots
print(roc1_plot)
print(roc2_plot)

library(car)

# Create the leverage plot
influencePlot(m2,
              main = "Leverage Plot for Model",
              sub = "Model Diagnostics: Residuals vs Hat-Values",
              col = c("blue", "red"), # Set colors for points
              xlab = "Hat-Values",   # X-axis label
              ylab = "Studentized Residuals" # Y-axis label
)

vif(m1)

vif(m2)
```

D. Dataset

Adult-NHIS-2023 Zip File:
https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2023/adult23csv.zip

Cleaned data: nhis1.csv