Data Analysis on Adult Mental Health in the US

Yu-Hsin Lee, Dr Sergey Kushnarev (Supervisor)

November 2024

1 Introduction

Mental illness is a major problem all over the world. In the US, mental health severity is often categorized as Any Mental Illness (AMI) and Serious Mental Illness (SMI), the latter being a subset of the former. It is estimated that more than one in five adults in the US live with AMI [5], a worrying trend. Mental health has intrinsic and instrumental value that enables us to cope with stress, realize our abilities and contribute to society [8]. In this project, we attempt to identify the most significant variables that affect adult mental health by finding the best model through statistical analysis. In doing so, we hope to highlight the adult groups with the highest vulnerability to mental health issues, so that assistance and preventive measures can be more readily promoted and advocated to these groups.

2 The Data

We examine the data from the 2022 National Survey on Drug Use and Health (NSDUH) provided by the Substance Abuse and Mental Health Services Administration (SAMHSA) [6]. NSDUH is an annual survey of the civilian, non-institutionalized population of the United States aged 12 years or older. While it is primarily a source of statistical information on the use of drugs in the population, it also provides valuable data on the demographics, mental health issues and mental health treatment among the general public. In this project, we analyze the influence of the background and drug use of an adult on their mental health. For predictors, we look at variables such as age, sexual orientation, gender, types and frequencies of drug use and many more (see appendix for variables); for response variables, we look at mainly *SMIPPPY* then *MICATPY* (as an extension), two recoded mental health indices provided in the dataset: the former is a predicted probability for SMI between 0 and 1; the latter is a predicted categorical indicator ranging from 0 to 3 in order of increasing mental health severity, with 3 representing SMI.

3 Methodology

This project is primarily an application of the methods and techniques of the course Applied Statistics and Data Analysis I (EN.553.413) together with the textbook Applied Linear

Statistical Models by Kutner et al. [2]. We first clean and perform exploratory data analysis, identifying issues with multicollinearity. Next, we fit an OLS model and evaluate diagnostic plots, determining potential issues to delve into. Then, we look into a combination of model selection techniques and remedial measures, comparing and analyzing them in order to find the best model for the data. Finally, we summarize the findings and identify groups that are most vulnerable to mental health issues.

4 Data Preparation and Exploratory Data Analysis

4.1 Data Preparation

Majority of the predictor variables we select are categorical, usually coded within the range of 10, with invalid (or less useful) data coded with numbers such as 85 - BAD DATA, 94 - DON'T KNOW, 97 - REFUSED, etc. Invalid data are a small portion of the data, so they are removed. For frequency of drug use, the value is coded between 1 and 30, with 91 representing NEVER USED ALCOHOL and 93 representing DID NOT USE ALCOHOL IN THE PAST MONTH. We recode DID NOT USE ALCOHOL IN THE PAST MONTH to 0 and NEVER USED ALCOHOL to -1. We also drop rows with *nan*. Ultimately, the total number of data points is reduced from 50510 to 37846, and 39 predictor variables are selected. Categorical variables are converted to factor variables in R, while numerical variables are standardized accordingly.

Since $SMIPPPY$ is a probability between 0 and 1, we perform logit transform on it to create a new response variable:

$$Y_{SMIPPPY} = \log \frac{SMIPPPY}{1 - SMIPPPY}$$

$Y_{SMIPPPY}$ is the main response variable in this project.

4.2 Exploratory Data Analysis

By removing *nan*, we obtain a fill rate of 100% for all predictor variables. We plot a covariance heat map for every pair of predictor variables (Figure 1). We also examine the adjusted generalized variance inflation factor (adjusted GVIF) from a simple OLS model. The GVIF is the Variance Inflation Factor corrected by the number of degrees of freedom of the predictor variable, and since each of the predictors have different number of factor levels, the adjusted GVIF is used to make GVIF comparable across different number of parameters. [1] The adjusted GVIF is given by $\text{GVIF}^{1/(2df)}$. Since the adjusted GVIF of all variables are lower than 2 (Figure 2), we assume little multicollinearity is present.

5 Diagnostics

Assuming uncorrelated, zero mean and homoscedastic errors, the Gauss-Markov theorem states that the Ordinary Least Squares (OLS) estimator is the best linear unbiased estima-

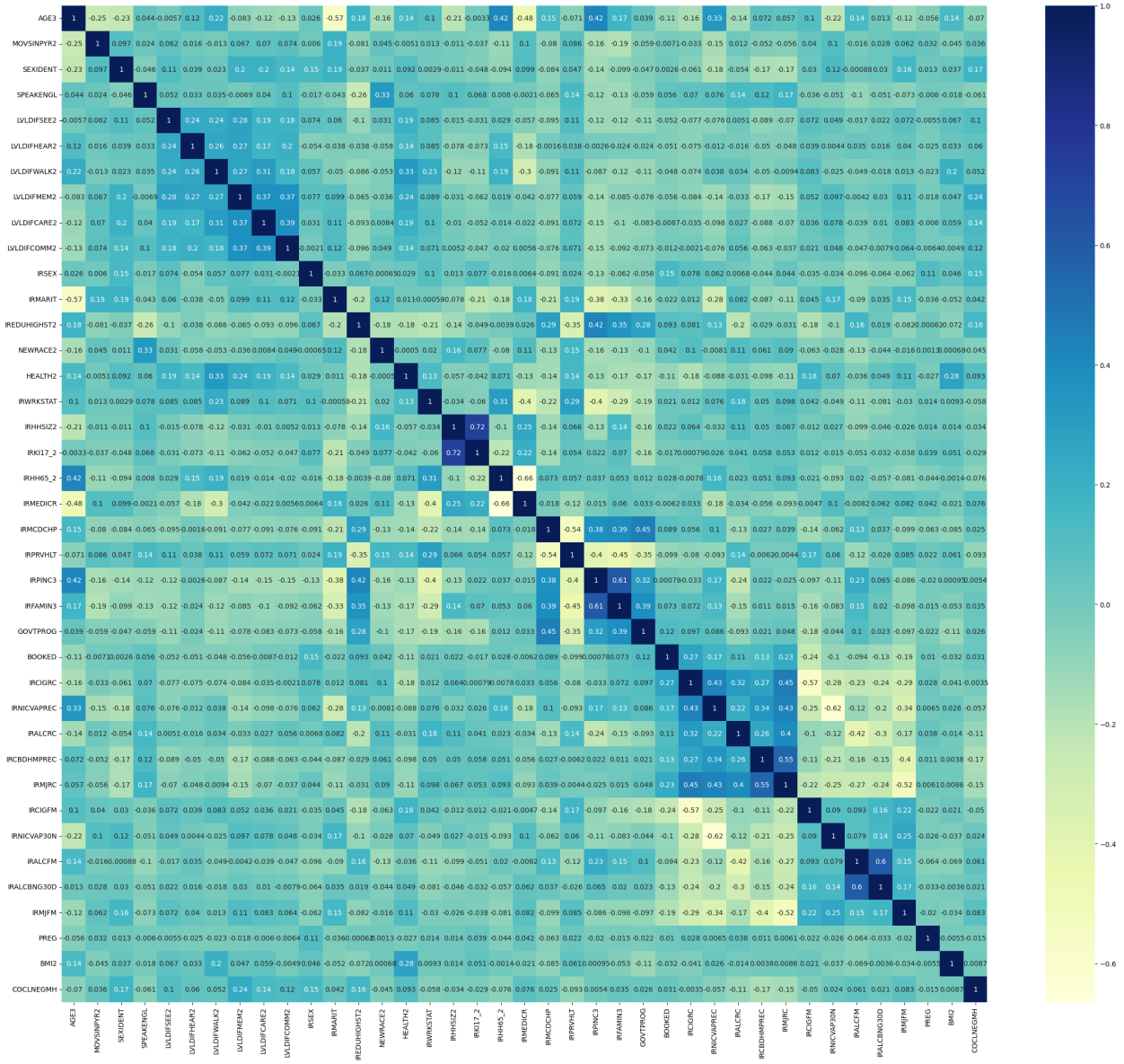

Figure 1: Covariance Heat Map

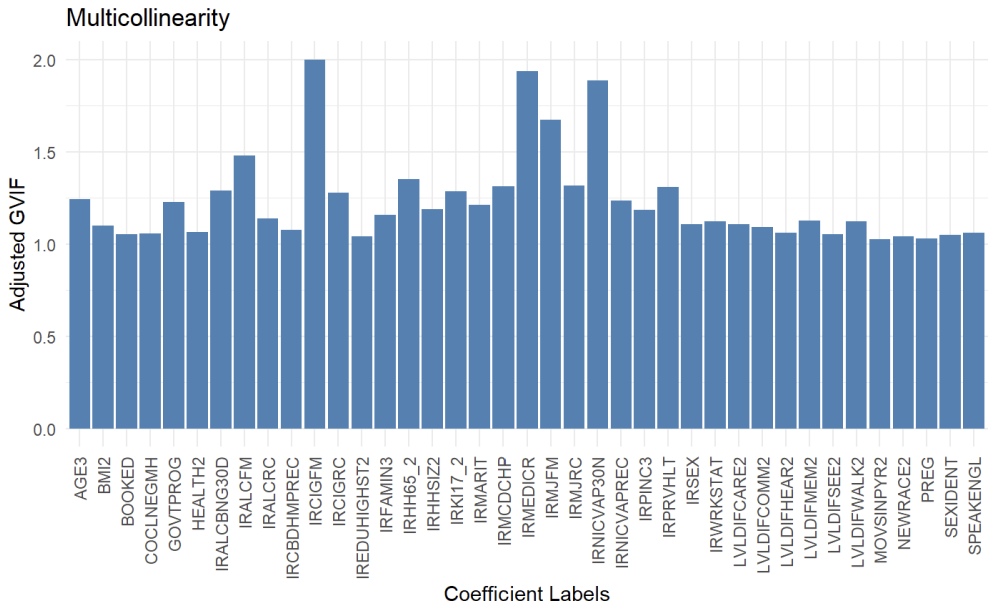

Figure 2: GVIF plot

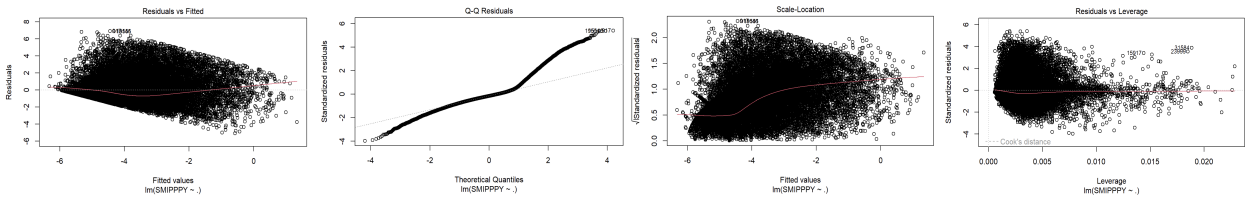

Figure 3: OLS Diagnostics

tor with the smallest variance. We first fit a simple OLS model in R and plot the diagnostics.

Figure 3 shows Residuals vs Fitted, Q-Q, Scale-Location and Residuals vs Leverage plots. Heteroscedasticity can be observed from the first - lower residual variance in the extreme values of Fitted. In fact, the plot resembles a downward linear parallelogram between the residuals and fitted values. Non-normal errors with heavy left and right tails can also be observed.

Figure 4 is a plot of Studentized residuals vs Fitted values. Studentized residuals are more effective in detecting outliers and influential data points by taking into consideration the residuals may have substantially different variances. The formula is given by:

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

where the denominator is an estimator of the standard deviation of e_i, the i^{th} residual, with MSE the mean squared error, and h_{ii} the hat value located on the i^{th} row and column of the hat matrix. In the plot, even when studentized, the residuals do not have constant variance, nor is the Loess curve hovering close to zero at the extreme values of Fitted. Such a trend

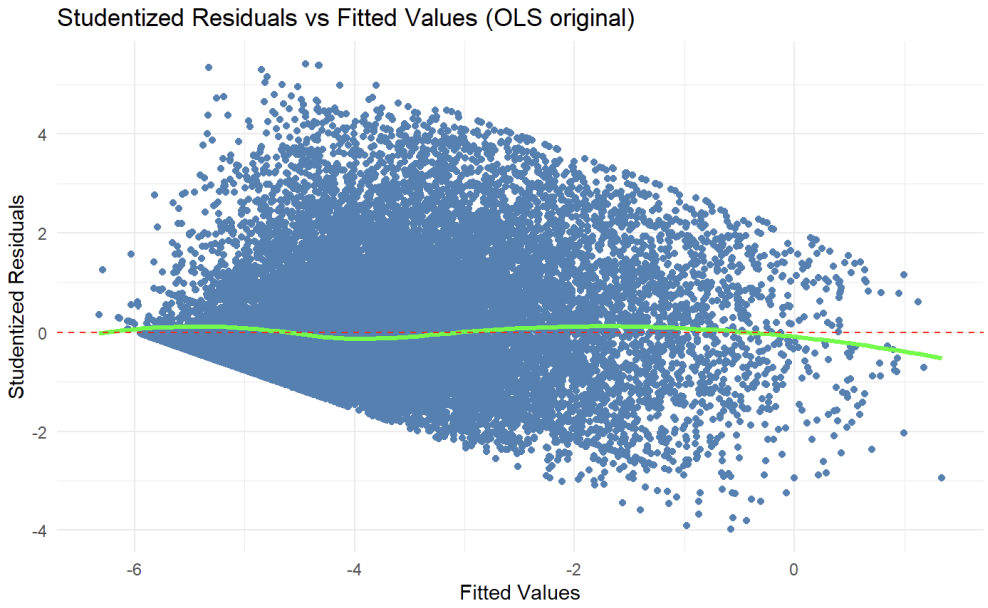

Figure 4: OLS Studentized Residuals

in the residuals may also suggest non-linearity.

6 Model Selection

With so many predictor variables, it is natural to try to reduce the dimensionality of the problem. The presence of more inter-correlated variables may increase sampling variation of regression coefficients, detract from the model's descriptive abilities, increase the problem of roundoff errors, and perhaps even worsen the model's predictive ability. Besides, regression models with fewer predictors would be easier to understand. [2] There are many ways to select the best subset of predictors, using criteria such as adjusted R^2, AIC, BIC, PRESS values, etc., but due to the large dimensionality of data, we use an automatic search for ease of selection.

6.1 Backward Elimination Regression

Backward Elimination begins with the model containing all predictors, then eliminates predictors using various criterion. Here, we use the *stepAIC* function from the *MASS* package which eliminates predictors using Akaike's Information Criterion (AIC), resulting in a final model with 33 predictors, though AIC is only reduced from 124198.1 to 124182.9. The coefficients are shown in figure 8.

6.2 LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) is another technique that performs variable selection and regularization. It constrains the magnitude of OLS coefficients

using L1 norm, i.e. its objective function is, for some regularization constant λ:

$$\min_{\beta} \left\{ \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \right\}$$

With the L1 norm constraint, LASSO is able to shrink coefficients to zero, allowing for variable selection. We perform weighted LASSO, i.e., LASSO combined with weights (see section 7.1) using the *glmnet* package, which uses cross validation to find the optimal λ. However, we obtain $\lambda = 0.00342$, a close-to-zero value for the regularization constant, suggesting an almost OLS-like fit; hence, LASSO may not be the most appropriate for this problem.

7 Remedial Measures

With non-constant variance in our residuals, we attempt to remedy it with methods such as Weighted Least Squares (WLS) and its extended version - Iteratively Reweighted Least Squares (IRLS), or we can fit robust regression models that do not assume constant variance in errors, including methods such as Least Absolute Deviations (LAD), and the robust regression version of IRLS. We also look into potential transformations to tackle the linear trend in the residuals.

7.1 Weighted Least Squares

Weighted Least Squares is often used to remedy the problem of heteroscedasticity. Assuming the error variances $\varepsilon_i \sim \text{Normal}(0, \sigma_i^2)$ i.i.d, we can obtain the weight matrix by approximating the variances s_i by fitting the model of absolute residuals against fitted values. By approximating the weight as

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{pmatrix}$$

where $w_i = \frac{1}{s_i^2}$, we then obtain the WLS coefficients $b_{WLS} = (X^T W X)^{-1} X^T W Y$.

See Figure 5. The LOESS curve is now hovering close to zero throughout, illustrating the effects of WLS, but from the plot of studentized residuals, the residuals still have non-constant variance, suggesting that other factors may be in place. The coefficients are shown in figure 7.

7.2 Iteratively Re-weighted Least Squares

As the name suggests, this method is an iterated version of WLS. Because the variances of the errors are unknown, the weights for WLS are merely estimates. As such, the variances of the errors may not truly be constant after one iteration. With IRLS, we repeat the WLS process by estimating the error variances using the absolute residuals vs fitted values fit at each step,

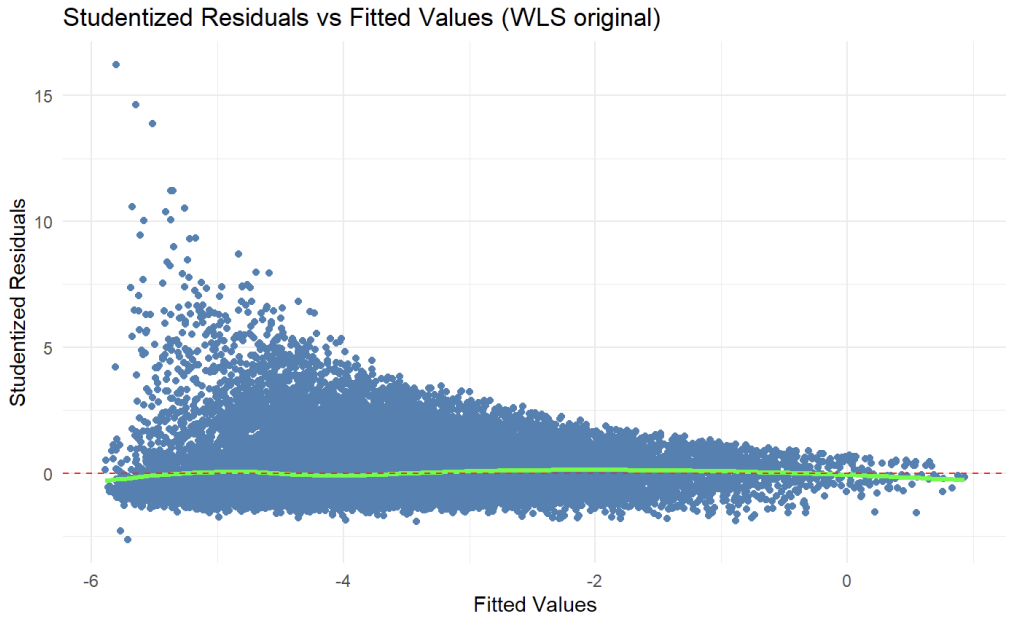

Figure 5: WLS Studentized Residuals

until the change in coefficient magnitudes fall below a certain threshold; here, we set this change to be the maximum relative change of all coefficients, $\max((\beta_{\text{new}} - \beta_{\text{old}})/\beta_{\text{old}})$, with a threshold of 0.001. From the textbook, the mean squared error MSE_w can be estimated by

$$MSE_w = \frac{\sum w_i e_i^2}{n - p}$$

where n is the total number of observations and p is the number of parameters. This model achieved a MSE_W of 2.38801, close to 1, suggesting a good fit. [2] The coefficients are shown in figure 7.

7.3 Least Absolute Deviations

Robust regression does not assume homoscedasticity, and dampens the influence of outlying cases, in order to provide a better fit for majority of the data points. In a way, robust regression methods also deal with non-constant variance because they reduce the effects of exceedingly large residuals on the model. One of such regression methods is the Least Absolute Deviations (LAD) method, otherwise known as the minimum L_1 norm regression. Its objective function is similar to OLS, except that it minimizes the L_1 norm rather than the L_2 norm, i.e.

$$\min \sum_{i=1}^{n} |Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)})|_1$$

Because of the L_1 norm, LAD regression places less emphasis on observations where the residuals are greater. We use the *lad* function from the *L1pack* library to fit a LAD model. The coefficients are shown in figure 7.

7.4 IRLS Robust Regression

Robust IRLS is another remedial measure for heteroscedasticity and outlying influential observations. Like IRLS, it reduces the influence by employing weights inversely related to the size of the residual, but this time according to weight functions: outliers with large residuals will be given less weight based on these functions. Huber or Bi-square function are most commonly used. Following the initialization of weights, for each iteration, we fit the WLS model and use the residuals to obtain revised weights via the weight function. This is repeated until convergence.

We use the rlm package in the *MASS* library, setting 100 max iterations first for Huber function, then 200 max iterations for Bi-square function, the latter being more discriminatory. The coefficients are shown in figure 7.

7.5 Transformations

Non-linearity in the residuals may warrant a transformation in the predictors or the response. By possibly linearizing the relationship between transformed variables, these methods may also help in variance stabilization.

7.5.1 Box-Cox Transformation

Box-Cox transformation is used for strictly positive responses and it chooses the transformation based on data. It transposes response $y \to g_\lambda(y)$ where

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

For this data, we first shift the response upwards by 6 to ensure positiveness, then we use the *boxcox* function from the *MASS* package to find the optimal λ via negative log-likelihood - the studentized residuals are shown in figure 6. Observe that the linear trend in residuals still persists, yet the LOESS curve is worsened.

7.5.2 Generalized Linear Models

From the original studentized residuals plot, figure 4, we observe it resembles lines with -1 gradient, which is common among discrete responses: for fixed $y = k$, residuals $\hat{\varepsilon} = y - \hat{y} = k - \hat{y}$, which has a slope of -1. By converting the response into a discrete one, we can find models that may explain this phenomenon. We first fit a Binomial GLM by transforming a new response using a threshold of 0.5, i.e. $y_{bin} = 1$ if $SMIPPPY > 0.5$ and $y_{bin} = 0$ if $SMIPPPY \leq 0.5$. This resulted in a very significant p-value of less than 2.2×10^{-16} using the Hosmer and Lemeshow goodness of fit test, suggesting a very poor fit. We also fit a Poisson model by using $y_{poi} = \text{round}(Y_{SMIPPPY} + 6)$. Similar methods can be explored more in-depth in the future.

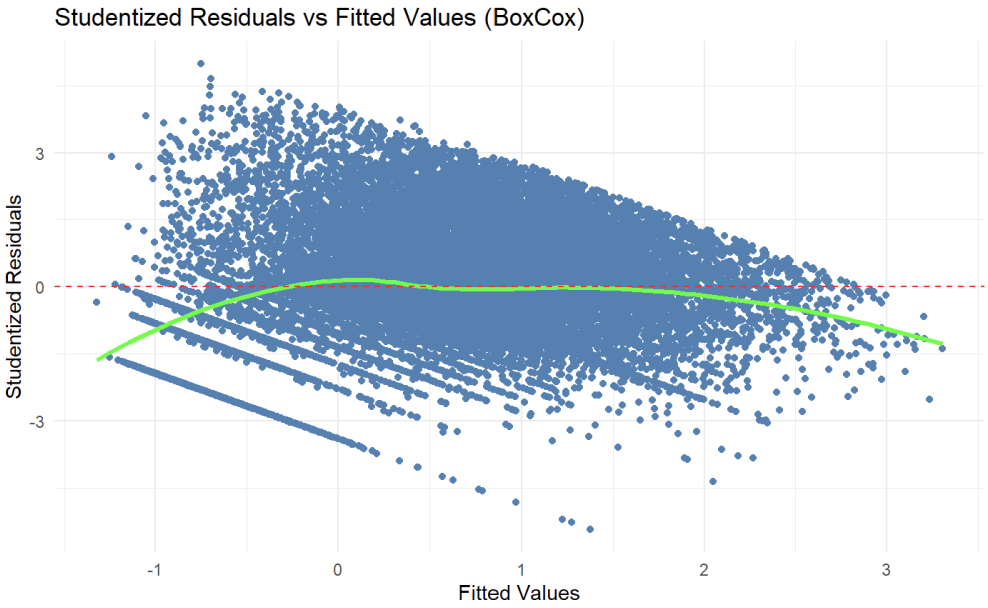

Figure 6: BoxCox Studentized Residuals

8 Interpretation

We combine the coefficients of OLS, WLS, IRLS, LAD and IRLS Robust Regression in figure 7. Observe that the coefficients of WLS, IRLS, LAD, IRLS Robust Regression agree with one another, amplifying and diminishing the same coefficients. LAD, in particular, minimized most of the less significant variables close to zero. Figure 8, comparing LAD and backward models, also shows that the more significant variables these models have selected are a subset of the backward selection model, hence further solidifying their importance. Using the *White* estimator, $S^2(b) = (X^TX)^{-1}(X^TS_0X)(X^TX)^{-1}$ where S_0 is a diagonal matrix consisting of the squares of ordinary least squares estimators of residuals, to estimate the variance of coefficients obtained from IRLS, we can test the significance of each coefficient obtained, and annotate on the plot with an asterisk if significant. [2] Referring back to figure 7, the number of asterisks indicates the level of significance of the coefficient, and 4 asterisks indicate that the p-value is less than 2.2×10^{-16} (very significant).

In this plot, intercept represents the baseline (or reference point), i.e. the population that belongs to the specific group of AGE34 + BOOKED1 + COCLNEGMH2 + ... (all of these factor levels are 'removed' from the list of parameters and combined into the intercept to ensure non-singularity of the predictor matrix). Coefficients represent the change in $Y_{SMIPPPY}$ (logit SMIPPPY) if only the specific variable is changed while keeping all others constant at baseline. An increase in logit SMIPPPY implies an increase in SMIPPPY, i.e., a higher probability of SMI.

We observe that the significant variables include **age**, sex, **sexual orientation**, race, difficulties in sight, **communication**, **self-care** and **memory/concentration**, frequency of moving, **COVID aftermath** and **health** (bolded are the most significant).

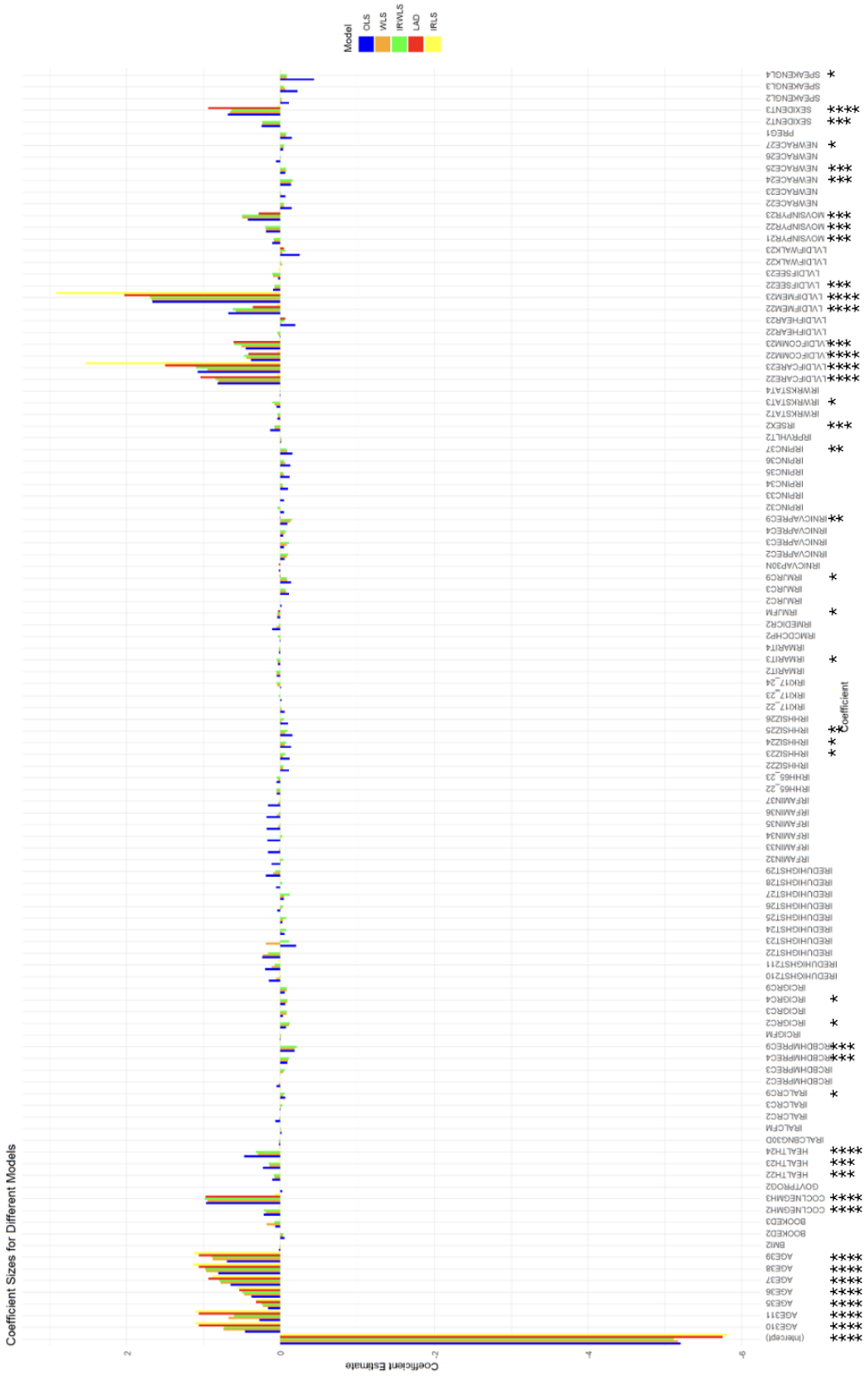

Figure 7: Coefficients for Different Models

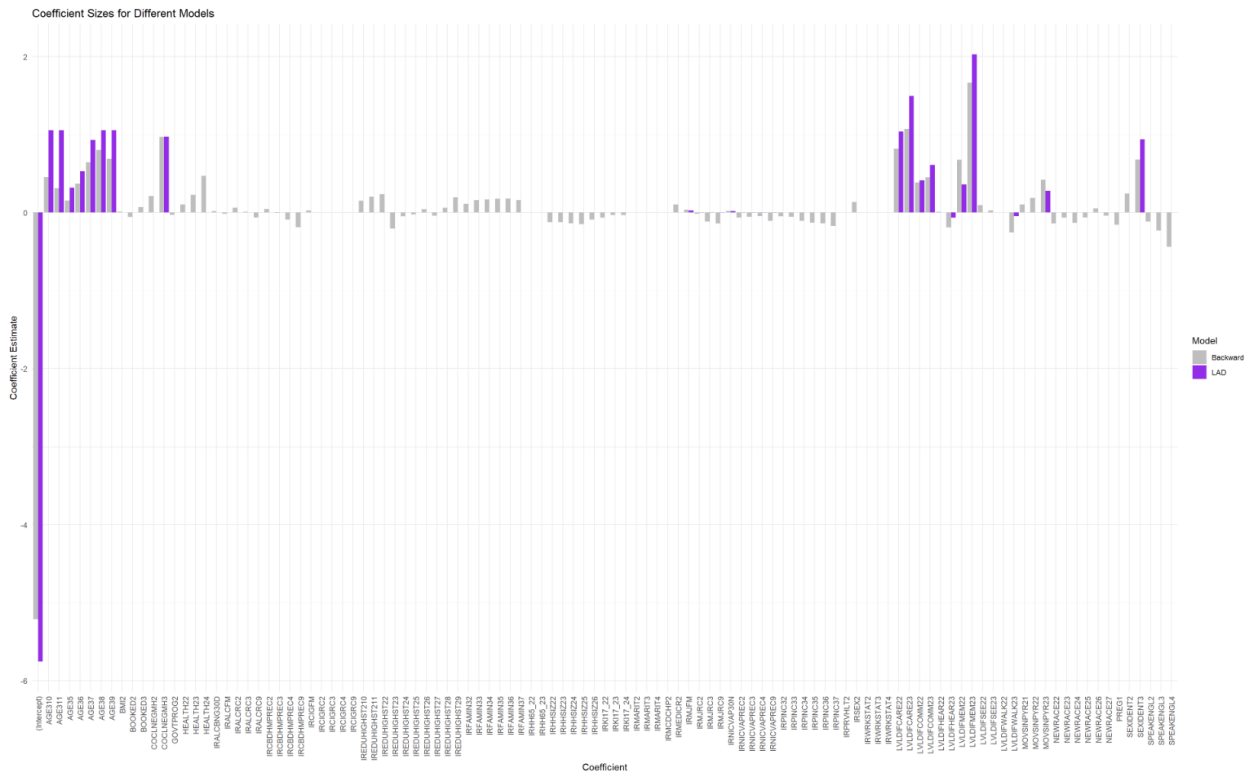

Figure 8: Coefficients for LAD and Backward Model

9 Conclusion

Most vulnerable groups include all age groups, people with impairments in communication, self-care and memory/concentration (all levels of severity), gay and bisexual people, people who have POOR health (out of EXCELLENT, VERY GOOD, GOOD, POOR) or were negatively affected by COVID (all levels of severity). Interestingly, higher age groups incur greater change, as shown in figure by the size of the coefficients (but keep in mind the point of reference, i.e., having all other variables at baseline value); there may be a few explanations for this observation. For instance, there is a much larger proportion of non-heterosexual individuals in the younger generations, e.g. 23.2% of people aged 21-23 are gay or bisexual compared to 2.5% of people aged over 65. COVID-19 also seems to have affected younger individuals more negatively: 65.0% of people aged 21-23 indicated negative aftermath vs 57.3% of those aged over 65. Furthermore, there is a lot of room for improvement in terms of model fitting for such high-dimensional data, so the results may not be completely reflective of each variable's influence. What is important is that these highly vulnerable groups, which are also evidenced by other sources [3] [7], are highlighted and that we promote and extend existing help available more actively to these groups.

Special thanks to Dr Sergey Kushnarev for supervising the project, coming up with infinite ideas, explaining various statistical methods and making the experience very enjoyable :).

10 Extensions and Future Possibilities

As mentioned, *MICATPY* is a categorical response variable consisting of integers between 0 and 3 inclusive, in ascending order of mental health severity. Hence, we can consider an ordinal regression model for this problem. We fit an ordinal logistic regression using the *polr* function from the *MASS* package, giving a model with a McFadden's pseudo R^2 value of 0.221, suggesting an extremely good fit (between 0.2 and 0.4) [4]. Predictions on the dataset outputs an accuracy of 0.746, compared to a baseline of 0.721 (by the majority population with $MICATPY = 0$). Accuracy can be improved, and more relevant scores such as F-score, in particular type 1 and type 2 errors, can be further looked into.

10.1 Future Possibilities

A list of potential future extensions: interaction terms between variables, polynomials for numerical predictors, other variables in the dataset (out of the 2000+ variables) may be significant. The problem can also be extended to Machine Learning, with models such as random forest regressor models and neural networks.

11 Appendix

PREDICTORS in categories:

Age: AGE3
Gender/Sexuality: IRSEX, SEXIDENT
Physical/Mental Impairment: LVLDIFSEE2, LVLDIFHEAR2, LVLDIFWALK2, LVLDIFMEM2, LVLDIFCARE2, LVLDIFCOMM2
Education: IREDUHIGHST2
Race: NEWRACE2
Family: IRMARIT, IRHHSIZ2, IRKI17_2, IRHH65_2
Employment/Income: IRWRKSTAT, IRPINC3, IRFAMIN3, GOVTPROG
Health: HEALTH2, PREG, BMI2, COCLNEGMH
Healthcare: IRMEDICR, IRMCDCHP, IRPRVHLT
Drug Use: IRCIGRC, IRNICVAPREC, IRALCRC, IRCBDHMPREC, IRMJRC, IRCIGFM, IRNICVAP30N, IRALCFM, IRALCBNG30D, IRMJFM
Other: MOVSINPYR2, SPEAKENGL, BOOKED

RESPONSES:

SMIPPPY, MICATPY

References

[1] John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

[2] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, New York, NY, 5th edition, 2004.

[3] Cátia Laranjeira and Alexandra Querido. Mental health promotion and illness prevention in vulnerable populations. *Healthcare (Basel)*, 12(5):554, Feb 2024.

[4] Jordan J. Louviere, David A. Hensher, and Joffre Dan. Swait. *Stated choice methods : analysis and applications / Jordan J. Louviere, David A. Hensher, Joffre D. Swait ; (with a contribution by Wiktor Adamowicz).* Cambridge University Press, Cambridge, U.K. ;, 2000.

[5] National Institute of Mental Health. Mental illness, 2023. Accessed: 2024-11-11.

[6] Substance Abuse and Mental Health Services Administration. Nsduh: National survey on drug use and health, 2023. Accessed: 2024-11-11.

[7] World Health Organization. Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide, March 2022. Accessed: 2024-11-25.

[8] World Health Organization. Mental health: Strengthening our response, 2024. Accessed: 2024-11-28.