Student Mental Health: A Statistical Exploration of Stress and Well-Being

Vajinder Kaur
Master of Science Statistical Practice
Boston University
Professor : Dr. Fotios Kokkotos
Github

8 December 2024

Abstract

India, despite being the most populated country, is geographically smaller compared to the largest nations. While the country is rapidly evolving, some issues persist, particularly in the domain of mental health. Mental health has long been a sensitive topic, often considered taboo, leading to a lack of open discussion and awareness. Depression, one of the most prevalent mental health concerns, is frequently misunderstood and tagged as a mental illness. This results in delayed identification and treatment for those affected, especially students in India. This study aims to address these challenges by investigating the factors contributing to depression among students and exploring methods to identify it in its early stages. By shedding light on this critical issue, the research hopes to contribute to a more informed understanding and proactive approach to mental health among the youth in India.

1 Introduction

This study is conducted using a student depression dataset that contains data aimed at analyzing, understanding, and predicting depression among students. It includes features such as demographic (Age, Gender, City), academic (CGPA, Degree, Academic Pressure, Study satisfaction), lifestyle habits (Sleep Duration, Dietary Habits, Study Hours), mental health history of Family, and Responses to Standardized Depression scales. Dataset contains responses from 27901 students and 18 variables. Each student is given a unique identifier (ID).

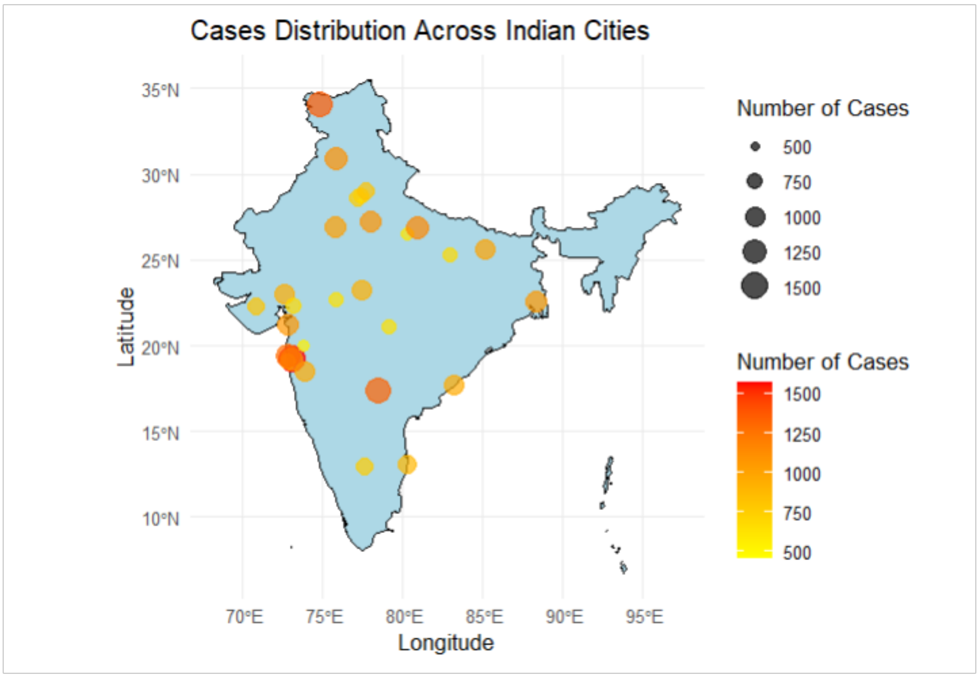


Figure 1: Distribution of Student Data Across Different Regions of India

In this dataset, cities like Hyderabad, Srinagar (Jammu and Kashmir), Kalyan (Maharashtra), Lucknow (Uttar Pradesh), Ludhiana (Punjab), Surat (Gujarat), Thane (Maharashtra), and Vasai-Virar (Maharashtra) have significantly higher numbers of cases compared to other cities due high popular density in these regions. However, the dataset still provides coverage across a wide range of regions, including the North, South, and West of India, representing some of the most densely populated states (A variable manually included after preprocessing data) in these regions. That being said, the dataset is not fully representative of East India. Historically, there has been limited data available from this region, though the reasons behind this disparity are complex and out of scope for this study.

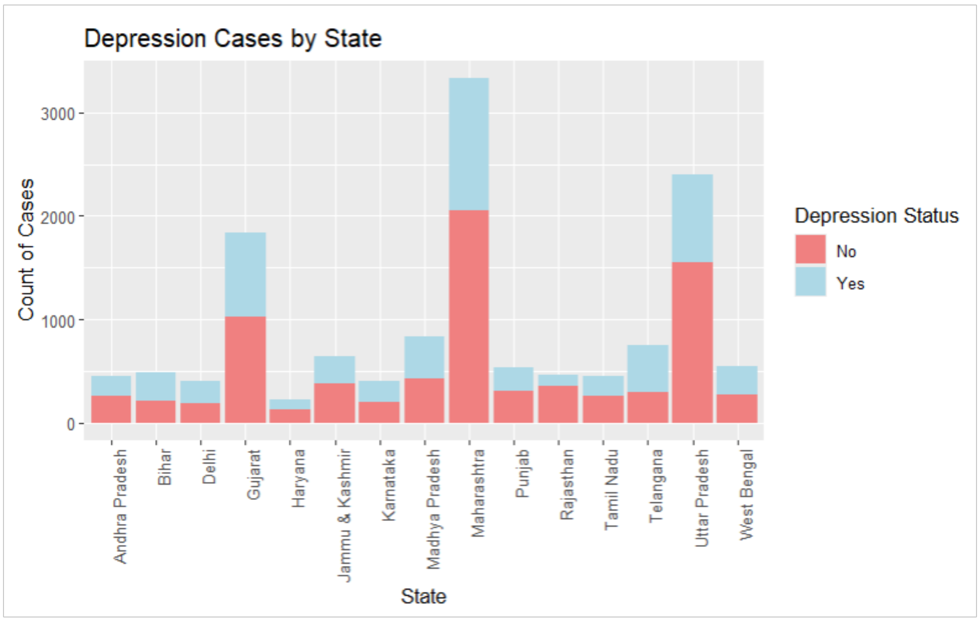


Figure 2: Distribution of Depression Cases Across Different States of India

With this background established, we now proceed to perform an exploratory data analysis (EDA) to further explore the dataset. Following that, we move on to the modeling phase, where the results are presented.

Methodology

The problem at hand is being addressed using logistic regression with a focus on modeling binary outcomes. The logistic regression model can be extended to incorporate varying degrees of pooling to account for hierarchical or grouped data. In this study, I will explore three approaches: complete pooling, no pooling, and partial pooling, each being compared for model fit and predictive performance. In addition, I will also try using probit if potential outliers are detected.

Logistic Regression Model

The general form of the logistic regression model is:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where:

- μ is the probability of success (i.e., the probability that the binary outcome is 1).
- $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients to be estimated.
- $X_1, X_2, \ldots, X_p$ are the predictor variables.

The logistic regression model applies the *logit* link function (in $glm()$ by default it is logit while using $family = binomial()$) to model the probability of success in a binary outcome. This ensures the predicted probabilities are always between 0 and 1.

Pooling Approaches

The data may be structured hierarchically (e.g., observations within groups), which can be modeled with varying levels of pooling.

Complete Pooling

In the complete pooling approach, all observations are treated as coming from a single, homogeneous group. The model does not differentiate between groups or subgroups and assumes a global model for all data.

The logistic regression model under complete pooling is: (Which is same as general Logistic Regression)

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

This model assumes that the same set of coefficients applies across all groups.

No Pooling

In the no pooling approach, we assume that each group has its own unique logistic regression model. This method does not share any information between groups and estimates separate coefficients for each group.

For a group-specific logistic regression model, the formulation is:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_{0,i} + \beta_{1,i} X_{1,i} + \cdots + \beta_{p,i} X_{p,i}$$

Where μ_i is the probability of success for observation i in group i, and the coefficients $\beta_{0,i}, \beta_{1,i}, \ldots, \beta_{p,i}$ are estimated separately for each group. This approach may suffer from overfitting, especially if the number of observations per group is small.

Partial Pooling

Partial pooling is modeled using hierarchical models(in this case I will be using $glmer()$), which share information across groups while allowing for group-specific variations. The model assumes that the group-specific parameters are drawn from a common distribution, typically a normal distribution. This approach helps to balance between the complete and no pooling approaches by borrowing strength from the entire dataset.

In a hierarchical logistic regression model, the group-specific parameters are modeled as:

$$\beta_{0,i} \sim N(\mu_0, \sigma_0^2), \quad \beta_{1,i} \sim N(\mu_1, \sigma_1^2), \quad \ldots, \quad \beta_{p,i} \sim N(\mu_p, \sigma_p^2)$$

Where $\mu_0, \mu_1, \ldots, \mu_p$ are the global means for each coefficient, and $\sigma_0, \sigma_1, \ldots, \sigma_p$ are the standard deviations that represent the variability between groups. The model allows for group-specific deviations from the global mean, but also ensures that extreme values are regularized.

The model is estimated using the `glmer` function in the `lme4` package, which fits generalized linear mixed-effects models.

Model Evaluation

To evaluate the model performance under each pooling strategy, we can use the Akaike Information Criterion (AIC). The model with the lowest AIC is considered the best fit for the data. The AIC provides a balance between model fit and complexity, with a penalty for adding more parameters. If overdispersion or a lack of fit is observed, alternative models such as quasi-binomial regression or negative binomial regression could be considered. In addition to AIC, I have used Confusion Matrix to compare accuracy of different models.

Quasi-binomial Regression

If overdispersion is detected, I may consider adjusting the model using quasi-binomial regression. The quasi-binomial model is similar to the logistic regression model but allows for overdispersion by introducing a dispersion parameter, ϕ. The likelihood function is adjusted to account for the dispersion.

The quasi-binomial model is given by:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where the variance for the quasi-binomial is:

$$\text{Var}(Y) = \mu(1-\mu)\cdot\phi$$

Here, $\phi > 1$ represents overdispersion, and $\phi = 1$ corresponds to the standard binomial variance.

2 Preprocessing Data

Source of dataset : Student Depression Dataset
The dataset initially contained only three missing values (NAs) in the Financial.Pressure column, which were removed during preprocessing. Several character categorical variables were transformed (encoded) into numeric factors for analysis, including:

- **Gender**: Male = 0, Female = 1
- **Sleep Duration**: Less than 5 hours = 1, 5-6 hours = 2, 7-8 hours = 3, More than 8 hours = 4
- **Have you ever had suicidal thoughts?**: Yes = 1, No = 0
- **Family History of Mental Illness**: Yes = 1, No = 0
- **Dietary Habits**: Healthy = 1, Unhealthy = 2, Moderate = 3
- **New Degree**: Graduated = 1, Post Graduate = 2, Higher Secondary = 3 (converted from the original Degree column)

In addition, certain observations were removed due to extremely low counts (ranging from 5-10). These included entries such as 'Others' in Dietary Habits and Degree, age groups under 30, and categories like "No Academic Pressure" and "No Study Satisfaction." The **Work.Pressure** and **Profession** columns were also excluded as they contained only a single value. CGPA is the only continuous variables in this data as Age was converted into factor due to limited (13) unique values. States column is introduced for state by state analysis. It was formed using the City column in the dataset.

3 EDA

Firstly, We want to make sure some categories aren't underrepresented in our data by Age, Education Level and Dietary Habits. An outlier was removed after being found by Box Plot of CGPA. As Fig.3, and Fig.6 show, We have no skewed data for Age as we already removed data above 30. We have reasonable cases per category for dietary habits.

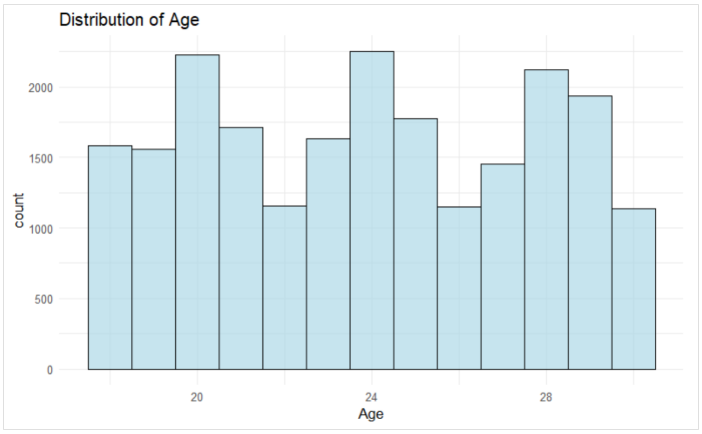


Figure 3: Age Distribution

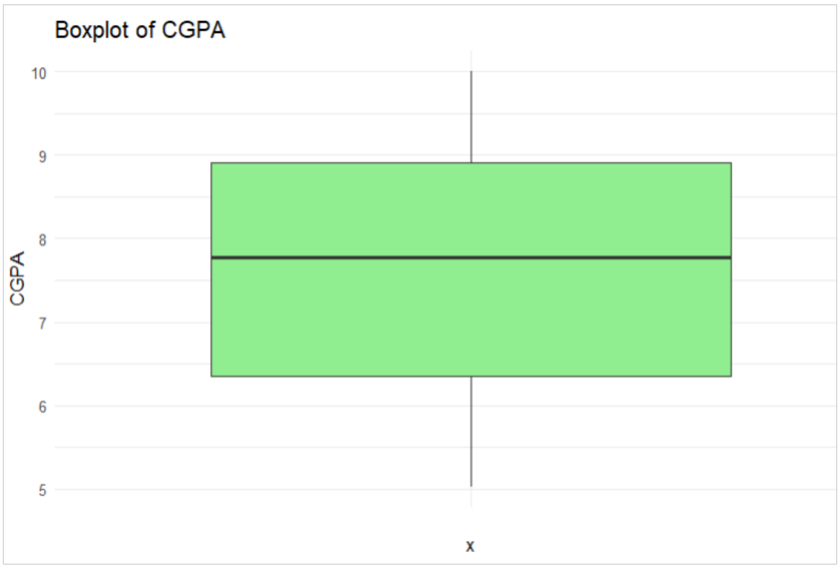


Figure 4: Distribution of CGPA

In Fig.5, we have more cases for graduated as compare to post graduate and higher secondary. It could be due to former being less in population overall than graduates in India and latter being very less aware about mental health. This can be supported by other papers.

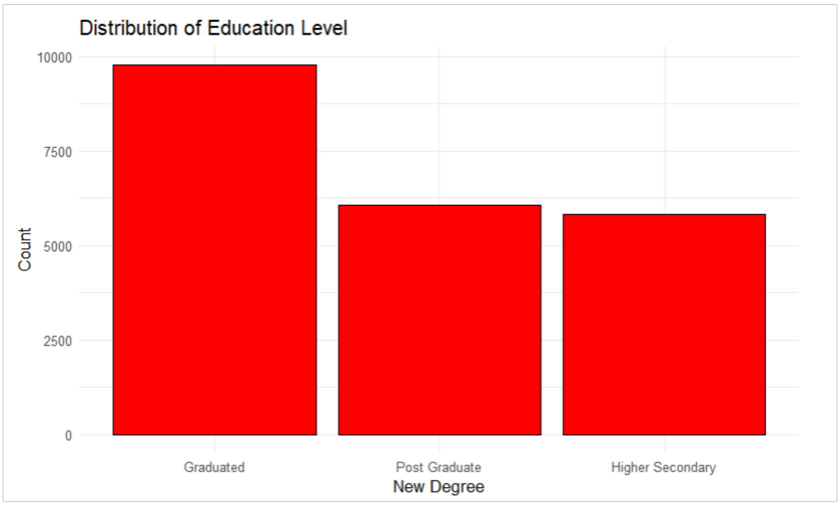


Figure 5: Distribution of Education

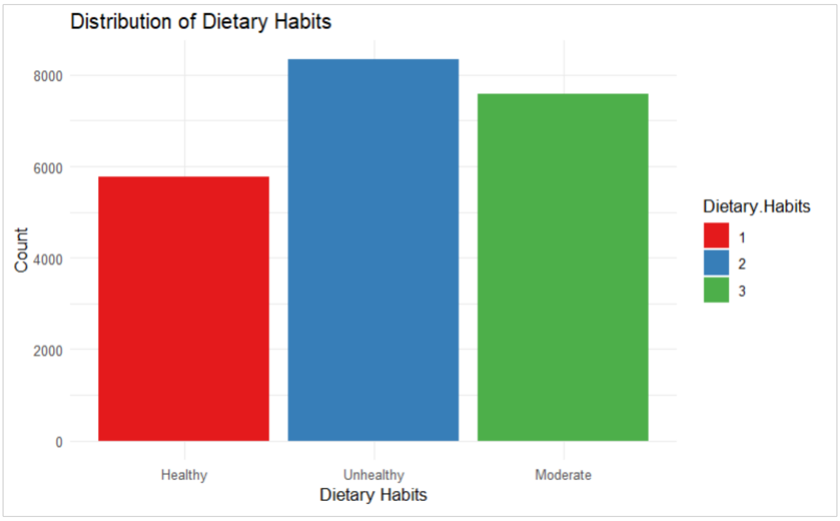


Figure 6: Distribution of Dietary Habits

The following are the results of the Pearson's Chi-squared tests for the relationships between various variables in the dataset:
(Have.you.ever.had.suicidal.thoughts.. is referred as Suicidal Thoughts for ease.)

Test	**Chi-Squared P-Value**
Depression vs Gender	0.08433
Depression vs Study Satisfaction	$< 2.2 \times 10^{-16}$
Depression vs Academic Pressure	$< 2.2 \times 10^{-16}$
Depression vs Suicidal Thoughts	$< 2.2 \times 10^{-16}$
Depression vs Family.History.of.Mental.Illness	$< 2.2 \times 10^{-16}$
Depression vs Sleep Duration	$< 2.2 \times 10^{-16}$
Suicidal Thoughts vs Family History of Mental Illness	1.943e-05
Academic Pressure vs Study Satisfaction	$< 2.2 \times 10^{-16}$
Dietary.Habits vs Family History of Mental Illness	0.4338
Work.Study.Hours vs Family.History.of.Mental.Illness	0.359
Sleep.Duration vs Study.Satisfaction	0.06919

The chi-square test indicates that the relationships between all variables and Depression(Response), except for Gender, are statistically significant. To my surprise, Gender does not exhibit a strong association with the other variables, which challenges the common belief that females in India are more likely to suffer from mental illness due to factors such as high crime rates, societal pressures, and issues related to freedom and lifestyle. This finding invites further investigation into the complexities of gender and mental health in India, suggesting that additional factors may play a more prominent role than gender alone.

Shedding a light on relationship between all variables (in order to choose predictor for our null model), only variables that didn't show significant relationship between themselves were Study.Satisfaction, Sleep.Duration and Family.History.of.Mental.Illness. So, we chose these three variables for our Null model. I used T test to check the relationship between Depression and CGPA which was found statistical significant (p-value = 0.00234). I used linear regression to check the relationship between CGPA and other three predictors which were decided earlier. Having found not a strong relationship in that model, I proceeded to use CGPA as well in the Null Model which turned out to be a wrong decision as it was found it violates assumption of Linearity between Log odds of Response variable and predictor for Logistic

Regression.

4 Modeling

4.1 Model Selection

Since the response variable (Depression) is binary, logistic regression is the most appropriate model for this study. Linear regression is typically used when the response variable is continuous, while Poisson regression is suitable for count data. Other models such as Linear Mixed Models, Quasi-Poisson, and Negative Binomial regression are also not suitable for this scenario, as they are designed for specific types of data (e.g., hierarchical data, over-dispersed count data). Logistic regression, on the other hand, effectively handles the binary nature of the response variable, making it the ideal choice for this analysis.

4.2 Models

Having explored Logistic Regression with Complete Pooling, No Pooling, and Partial Pooling, I ultimately decided on Complete Pooling or the Base Logistic Regression. The reasons for this choice will be explained in the later sections.

4.2.1 Null/Baseline Model

Predictors: Study.Satisfaction, Sleep.Duration, Family.History.of.Mental.Illness
Interpretation: Equation for the Null Model is below:

$$\begin{aligned}\log\left(\frac{\text{P(Depression=1)}}{\text{P(Depression=0)}}\right) &= 1.23804 - 0.30187 \cdot \text{Satisfaction2} - 0.55677 \cdot \text{Satisfaction3} \\ &- 0.80080 \cdot \text{Satisfaction4} - 0.99322 \cdot \text{Satisfaction5} \\ &- 0.36596 \cdot \text{Sleep4} - 0.22995 \cdot \text{Sleep3} \\ &- 0.58930 \cdot \text{Sleep4} + 0.25224 \cdot \text{FamilyHistory1}\end{aligned}$$

At a glance, it is evident that the log odds of a student with no family history of mental history, Sleep satisfaction = 1 and sleep duration =1 being depressed is 1.23804. This means odds of student from this group being depressed is 3.45 higher. As the Study Satisfaction increases for a student, their odds of depression decreases. Similarly, log-odds of students with sleep duration 5 - 6 hours being depressed decreases by 0.37 when compared to the student with less than 5 hours of sleep. Odds continue to decrease as sleeping hours increase. Student with family history of mental illness is more likely to be depressed.

4.2.2 Selected Model

Predictors: Age, CGPAlevel, Study.Satisfaction, Academic.Pressure, Have.you.ever.had.suicidal.thoughts.. , Work.Study.Hours, Financial.Stress, Family.History.of.Mental.Illness, Dietary.Habits, Sleep.Duration

Why CGPAlevel ? Upon converting it into categorical variable, linearity assumption was no longer a problem and it was still correlated to Depression. I decided to use all of the correlated variables because they contribute substantially to my response variable as per literature review which will be explained better in Discussion Section.
Interpretation: For the model results presented, all variables included are significant. During the modeling process, variables such as Gender and Degree were removed, which, surprisingly, were not significant.

The intercept in this model does not have a meaningful real-world interpretation, as Age = 0 is not realistic. Surprisingly increase in age decreases the log odds of depression by 0.0929, whereas an increase in grades is associated with a higher likelihood of depression. Increased academic pressure is strongly associated with higher log odds of depression. Greater satisfaction with studies is linked to a decreased likelihood of depression as coefficients are increasingly negative.

If someone has ever had suicidal thoughts, their log odds of depression increase by 2.485, indicating a very high likelihood of depression. More study hours are associated with increased depression, as a one-unit increase in study/work hours raises the log odds of depression by 0.11. This effect is really small if compared to suicidal thoughts. Additionally, a one-unit increase in financial stress increases the log odds of depression by 0.56. Family history of mental illness is positively associated with depression. Level 2 and 3 Dietary habits which corresponds to Moderate and Unhealthy eating habits also increases the chances of having Depression.

Table 2: Summary of Logistic Regression Results for Predictors of Depression (Rounded Off)

Predictor	**Estimate**	**Standard Error**	**z-value**	**p-value**
Fixed Effects				
Intercept	-5.29	0.15	-35.28	$< 2 \times 10^{-16}$
Age19	-0.36	0.12	-3.11	0.00190 **
Age20	-0.45	0.11	-4.15	3.26×10^{-5} **
Age21	-0.35	0.11	-3.02	0.00254 **
Age22	-0.68	0.12	-5.55	2.82×10^{-8} ***
Age23	-0.54	0.11	-4.73	2.29×10^{-6} ***
Age24	-0.59	0.11	-5.57	2.52×10^{-8} ***
Age25	-0.91	0.11	-8.27	$< 2 \times 10^{-16}$ ***
Age26	-1.11	0.12	-9.25	$< 2 \times 10^{-16}$ ***
Age27	-0.96	0.12	-8.30	$< 2 \times 10^{-16}$ ***
Age28	-0.95	0.11	-8.97	$< 2 \times 10^{-16}$ ***
Age29	-0.92	0.11	-8.54	$< 2 \times 10^{-16}$ ***
Age30	-1.76	0.12	-14.25	$< 2 \times 10^{-16}$ ***
CGPA_levelB	0.10	0.06	1.67	0.09584 .
CGPA_levelA	0.25	0.06	4.41	1.05×10^{-5} ***
Study.Satisfaction2	-0.38	0.07	-5.63	1.81×10^{-8} ***
Study.Satisfaction3	-0.44	0.07	-6.46	1.03×10^{-10} ***
Study.Satisfaction4	-0.77	0.07	-11.70	$< 2 \times 10^{-16}$ ***
Study.Satisfaction5	-0.99	0.07	-13.92	$< 2 \times 10^{-16}$ ***
Academic.Pressure	0.83	0.02	49.05	$< 2 \times 10^{-16}$ ***
Have.you.ever.had.suicidal.thoughts	2.48	0.04	56.28	$< 2 \times 10^{-16}$ ***
Work.Study.Hours	0.12	0.01	20.95	$< 2 \times 10^{-16}$ ***
Financial.Stress	0.56	0.02	36.57	$< 2 \times 10^{-16}$ ***
Family.History.of.Mental.Illness	0.26	0.04	6.35	2.18×10^{-10} ***
Dietary.Habits2	1.12	0.05	21.20	$< 2 \times 10^{-16}$ ***
Dietary.Habits3	0.48	0.05	9.34	$< 2 \times 10^{-16}$ ***
Sleep.Duration2	-0.44	0.06	-7.43	1.07×10^{-13} ***
Sleep.Duration3	-0.39	0.06	-6.85	7.18×10^{-12} ***
Sleep.Duration4	-0.67	0.06	-11.27	$< 2 \times 10^{-16}$ ***

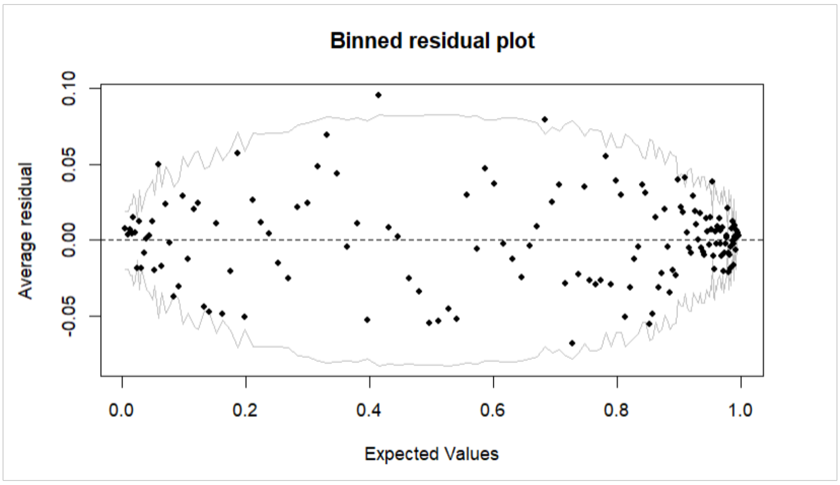


Figure 7: Binned Residual Plot

As all the residuals are scattered in randomly around the dotted line at 0, it shows model is a good fit. There are bunch of residuals in what looks like more dense group as expected values reaches 1, which tells there could be more improvement. I studied the potential outliers which could cause those three residuals outisde bin but having removed the unusual cases, there is nothing that can be done with the usual ones.

4.2.3 Assumptions Check

It was checked if all the assumption for the logistic regression were met. Below is a structured breakdown:

1. **Response Variable Suitability:** The binary nature of the response variable (*Depression: Yes/No*) aligns with the binomial family in GLM.

2. **Linearity of Predictors:**

   - As all of the variables are factors, I didn't really need to check the linearity.

3. **Multicollinearity:**

   - Variance Inflation Factors (VIFs) were all close to 1, indicating no significant multicollinearity among predictors.

4. **Adequate Observations Across Categories:**

   - The dataset has sufficient observations in each category, ensuring robust estimates and reliable model performance.

5. **Outlier Handling:**

   - Outliers were reviewed as logit is very sensitive to outliers, and only extreme deviations were removed after confirming that others represented valid data points. I tried probit as well in order to see if it improves or deals better with remaining outliers but it didn't perform any better than logit. In fact, it performed slightly worse than logit.

Variable	**GVIF**
Age	1.016630
CGPA	1.005112

Academic.Pressure	1.075409
Study.Satisfaction	1.012424
Dietary.Habits	1.018257
Have.you.ever.had.suicidal.thoughts..	1.079570
Work.Study.Hours	1.014447
Financial.Stress	1.038949
Family.History.of.Mental.Illness	1.002155

Table 4: GVIF values for the regression model (No Multicollinearity)

Overdispersion for GLMM was checked as well. As it was not overdispersed, I didn't need to go for Quasi-Binomial or Negative Binomial Regression.

4.2.4 Comparison between Models

As stated in the beginning of this section, three types of models were fitted for this study. All Logistic regression but with complete pooling, no pooling and partial pooling. As AIC, BIC and deviance shows, there is no significant difference in model performance. Even though GLMM is having lower AIC, but complexity of the model makes up for it. (Results from Probit and model with interactions are added in the Appendix)

Metric	**Null Model**	**Complete Pooling**	**No Pooling**	**Partial Pooling**
AIC	27633	15024	15024	15043.6
BIC	27704.41	15255.97	15235.01	15243.3
Deviance	27615	14966	14966	14993.6

Table 5: Model Comparison by AIC, BIC, and Deviance

Metric	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Accuracy	0.6428	0.85	0.85	0.85
95% CI	(0.6364, 0.6492)	(0.8452, 0.8547)	(0.8452, 0.8547)	(0.8452, 0.8547)
Kappa	0.0751	0.6711	0.6711	0.6713
Sensitivity	0.11751	0.7567	0.7567	0.7575
Specificity	0.94504	0.9037	0.9037	0.9032
Pos Pred Value	0.55154	0.8188	0.8188	0.8183
Neg Pred Value	0.65054	0.8659	0.8659	0.8662
Balanced Acc.	0.53127	0.8302	0.8302	0.8304

Table 6: Confusion Matrix Statistics for All Models

Overall Performance: GLMER (Partial Pooling) slightly outperforms both the Complete Pooling and No Pooling models in most metrics, including, specificity, precision, and balanced accuracy. Minor Differences: The differences between the models are relatively small, but GLMER (Partial Pooling) consistently shows slight improvements, especially in terms of Precision (Pos Pred Value) and Balanced Accuracy. In summary, while all three models perform similarly and have equal accuracy of 85

However, overall, there is not much difference in model performance. Effect of pooling isn't significant. If a simpler model is giving me nearly similar performance, why would I want to use complex one with little to no better performance?

Actually not even better but if we see model fit and AIC, Model with complete pooling is better than GLMM. It will save computational resources as well as time to use a simpler model. Easy interpretation of results is an another added advantage.

5 Discussion and Validation

Now comes the exciting part! Validation and discussing what we thought, what we got, and what is true.
Being from India myself, I had prior opinions about certain factors. Initially, I assumed that gender would play a significant role in depression, particularly thinking women in India might be more affected due to societal constraints. One survey study did get this result, where female students reported more depression than their counterparts. However, the results from our model showed that gender is not a significant factor. Upon reflection, this makes sense—while women face challenges like lack of independence and societal expectations, men are under immense pressure to succeed in careers, provide for their families, and conform to ideals of masculinity. These dual burdens make both genders equally vulnerable to depression, highlighting the need to address societal norms and provide mental health support inclusively. These statistics were surprising but nonetheless confirm that what our model found is accurate.

Another surprising finding was that as age increases, the likelihood of depression decreases. Recent studies suggest that Gen Z and the younger generation are more susceptible to depression than older generations. Thus, our study aligns with findings outside. However, some studies also point out that older people may simply be better at hiding their symptoms and dealing with them.

The relationship between academic pressure and depression has always been a focal point for such studies because of how impactful it could be to curb depression on a large scale if causal. Though not directly associated, studies so far indicate that academic pressure is one of the potential contributors to depression. This paper provides in-depth details about academic dynamics and depression, concluding a positive association between them. Overall, our study's results are quite compatible with other studies.

Another concerning part was that pooling did not make much difference across different models, which is due to dietary habits already explaining depression variability independently. Depression variability across different dietary habits is not significantly different, and each category had enough observations that complete pooling already performed well despite sparse data.

In the end, even though all of these factors are good determinant of depression in Indian Students, we still have many room for more exploration specially when it comes to Gender impact which is questionable.

References

[1] World Economic Forum. How India Sees Mental Health

[2] KFF. Young people are more likely than older adults to be experiencing symptoms of depression.

[3] PLOS ONE. Dietary Habits and Depression Symptoms.

[4] American Psychological Association. Link between food and mental health.

[5] Mind Voyage. India latest depression stats.

[6] ScienceDirect. Depression among Indian university students and its association with perceived university academic environment, living arrangements and personal issues.

[7] Wellbeing Port. Mental Health - Taboo in India.

[8] PubMed Central. Younger people are more vulnerable to stress, anxiety, and depression during COVID-19 pandemic: A global cross-sectional survey.

[9] Loyola eCommons. Why Depression is on the rise Amongst Millennials and Gen Z.

[10] Loyola eCommons. Is there a difference between elderly and younger patients with regard to the symptomatology and aetiology of depression?.

[11] ResearchGate. Academic Pressure and Depression.

[12] ScienceDirect. The association between academic pressure and adolescent mental health problems: A systematic review.

Appendix

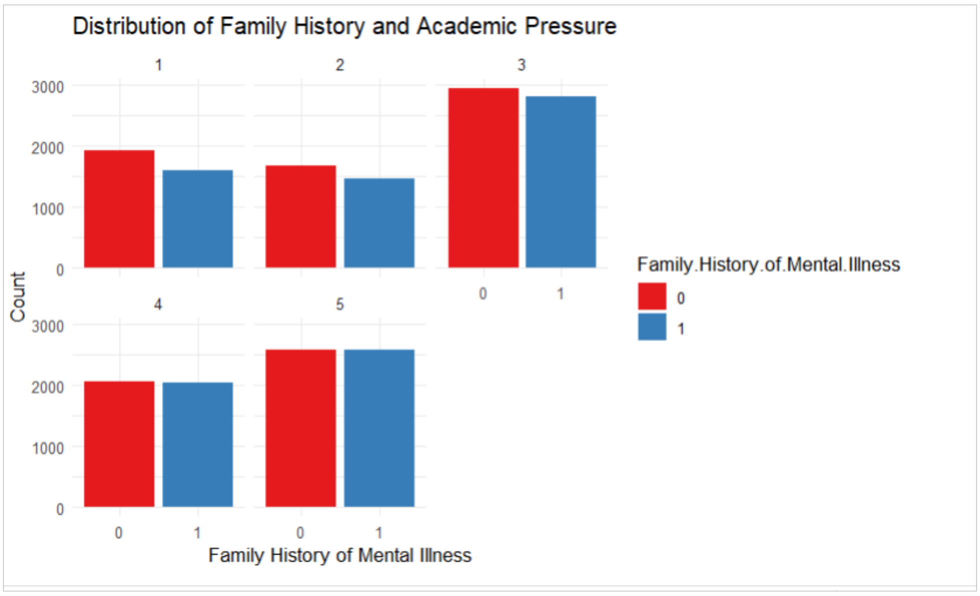


Figure 8: Distribution of Family History and Academic Pressure

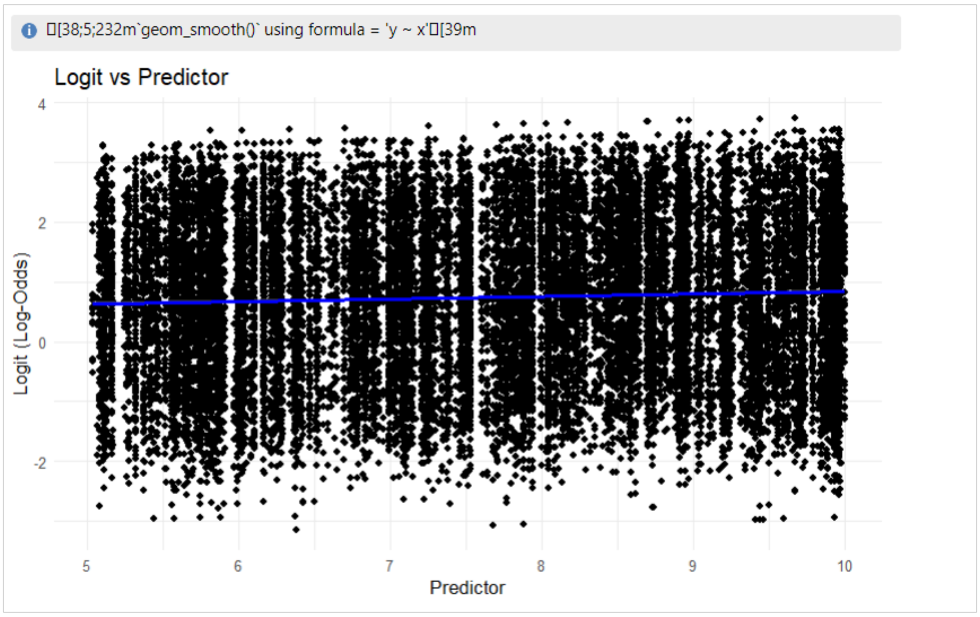

□[38;5;232m`geom_smooth()` using formula = 'y ~ x'□[39m

Figure 9: Linearity test

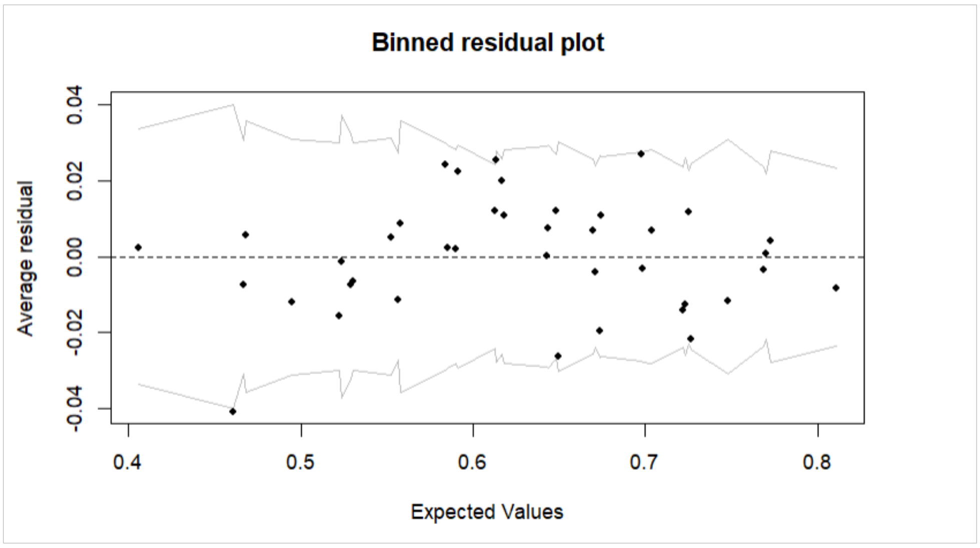


Figure 10: Residual Plot of Null Model

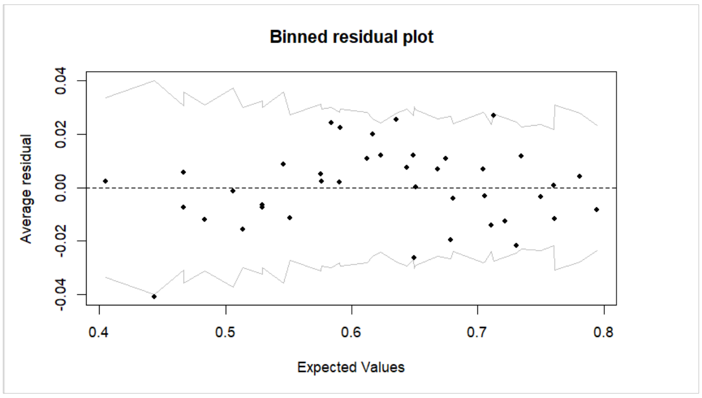


Figure 11: Residual plot of Null Model with Interactions

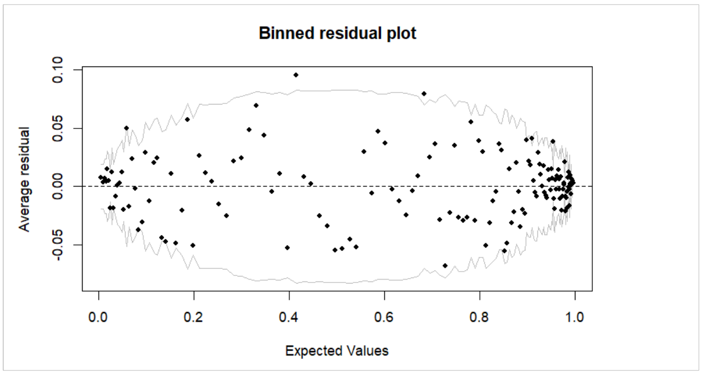


Figure 12: Residual plot of Probit Model

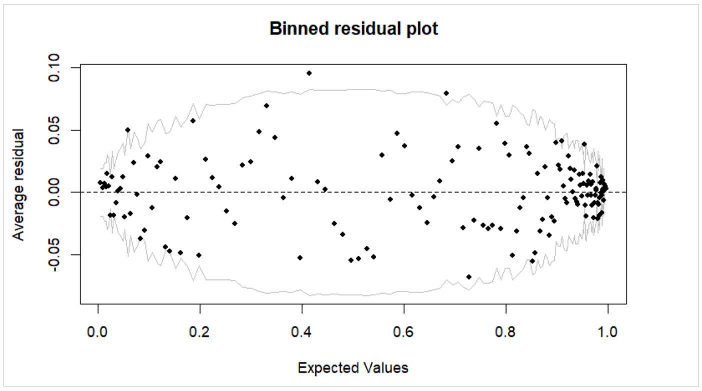


Figure 13: Residual plot of No Pooling Model

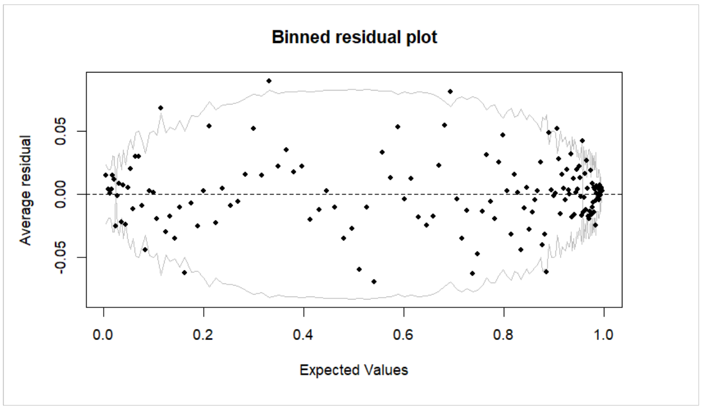


Figure 14: Residual plot of Partial Pooling

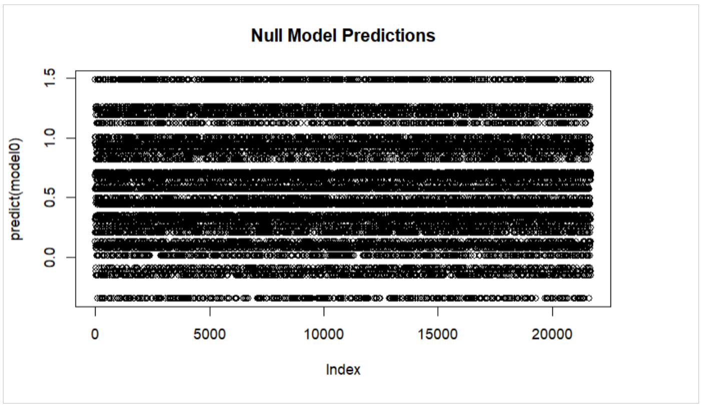


Figure 15: Prediction from Null Model

Deviance and AIC (Null Model)

- **Null deviance:** 28480 on 21695 degrees of freedom
- **Residual deviance:** 27615 on 21687 degrees of freedom
- **AIC:** 27633

Variable	**Coefficients**			
	Estimate	**Std. Error**	**z value**	**Pr(¿—z—)**
(Intercept)	1.23804	0.04424	27.982	$< 2 \times 10^{-16}$
Study.Satisfaction2	-0.30187	0.04735	-6.375	1.83×10^{-10}
Study.Satisfaction3	-0.55677	0.04705	-11.832	$< 2 \times 10^{-16}$
Study.Satisfaction4	-0.80080	0.04579	-17.488	$< 2 \times 10^{-16}$
Study.Satisfaction5	-0.99322	0.04917	-20.199	$< 2 \times 10^{-16}$
Sleep.Duration2	-0.36596	0.04071	-8.989	$< 2 \times 10^{-16}$
Sleep.Duration3	-0.22995	0.03916	-5.872	4.31×10^{-9}
Sleep.Duration4	-0.58930	0.04051	-14.546	$< 2 \times 10^{-16}$
Family.History.of.Mental.Illness1	0.25224	0.02884	8.746	$< 2 \times 10^{-16}$

Table 7: Coefficients from the Null Model for Depression.

Deviance and AIC(With Interactions)

- **Null deviance:** 28480 on 21695 degrees of freedom
- **Residual deviance:** 27608 on 21687 degrees of freedom
- **AIC:** 27626

Variable	**Coefficients**			
	Estimate	**Std. Error**	**z value**	**Pr(¿—z—)**
(Intercept)	1.298386	0.067976	19.101	$< 2 \times 10^{-16}$
Study.Satisfaction	-0.198978	0.019997	-9.951	$< 2 \times 10^{-16}$
Sleep.Duration2	-0.107842	0.100660	-1.071	0.28401
Sleep.Duration3	-0.008576	0.096787	-0.089	0.92939
Sleep.Duration4	-0.432951	0.098297	-4.404	1.06×10^{-5}
Family.History.of.Mental.Illness1	0.252207	0.028851	8.742	$< 2 \times 10^{-16}$
Study.Satisfaction:Sleep.Duration2	-0.084730	0.030228	-2.803	0.00506
Study.Satisfaction:Sleep.Duration3	-0.072338	0.028940	-2.500	0.01243
Study.Satisfaction:Sleep.Duration4	-0.050979	0.029691	-1.717	0.08598

Table 8: Coefficients from Null Model with Interaction for Depression.

Deviance and AIC(With Probit Link)

- **Null deviance:** 28480 on 21695 degrees of freedom
- **Residual deviance:** 14995 on 21667 degrees of freedom
- **AIC:** 15053

Variable	**Coefficients**			
	Estimate	**Std. Error**	**z value**	**Pr(¿—z—)**
(Intercept)	-2.993228	0.081627	-36.670	$< 2 \times 10^{-16}$
Age19	-0.213865	0.063828	-3.351	0.000806
Age20	-0.242182	0.059358	-4.080	4.50e-05
Age21	-0.193627	0.063459	-3.051	0.002279
Age22	-0.380621	0.067792	-5.615	1.97e-08
Age23	-0.296142	0.062932	-4.706	2.53e-06
Age24	-0.323350	0.058962	-5.484	4.16e-08
Age25	-0.514834	0.060885	-8.456	$< 2 \times 10^{-16}$
Age26	-0.602338	0.066983	-8.992	$< 2 \times 10^{-16}$
Age27	-0.536765	0.064024	-8.384	$< 2 \times 10^{-16}$
Age28	-0.528380	0.058928	-8.967	$< 2 \times 10^{-16}$
Age29	-0.503614	0.059873	-8.411	$< 2 \times 10^{-16}$
Age30	-0.973744	0.068219	-14.274	$< 2 \times 10^{-16}$
CGPA_levelB	0.056219	0.032664	1.721	0.085226
CGPA_levelA	0.143994	0.031476	4.575	4.77e-06
Study.Satisfaction2	-0.198987	0.037568	-5.297	1.18e-07
Study.Satisfaction3	-0.227456	0.037527	-6.061	1.35e-09
Study.Satisfaction4	-0.412597	0.036573	-11.281	$< 2 \times 10^{-16}$
Study.Satisfaction5	-0.536467	0.039457	-13.596	$< 2 \times 10^{-16}$
Academic.Pressure	0.464183	0.009164	50.652	$< 2 \times 10^{-16}$
Have.you.ever.had.suicidal.thoughts..1	1.414223	0.023991	58.949	$< 2 \times 10^{-16}$
Work.Study.Hours	0.067053	0.003150	21.288	$< 2 \times 10^{-16}$
Financial.Stress	0.314070	0.008392	37.424	$< 2 \times 10^{-16}$
Family.History.of.Mental.Illness1	0.147481	0.023181	6.362	1.99e-10
Dietary.Habits2	0.626919	0.029183	21.482	$< 2 \times 10^{-16}$
Dietary.Habits3	0.273823	0.028826	9.499	$< 2 \times 10^{-16}$
Sleep.Duration2	-0.245771	0.032605	-7.538	4.78e-14
Sleep.Duration3	-0.214439	0.031382	-6.833	8.31e-12
Sleep.Duration4	-0.369867	0.032879	-11.249	$< 2 \times 10^{-16}$

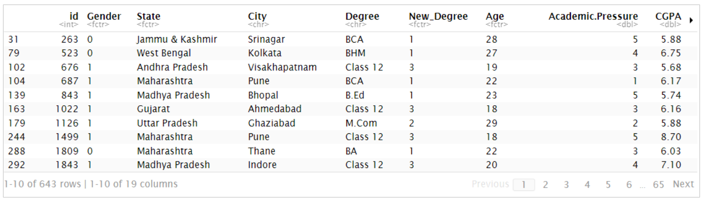
Table 9: Coefficients from the GLM for Depression with Probit Link Function.

Model Summary

- **Family:** binomial (logit)
- **AIC:** 15040.7
- **BIC:** 15264.2
- **logLik:** -7492.3
- **Deviance:** 14984.7
- **df.resid:** 21668

Variable	**Coefficients**			
	Estimate	**Std. Error**	**z value**	**Pr(¿—z—)**
(Intercept)	-4.757485	0.300099	-15.853	$< 2 \times 10^{-16}$
Age19	-0.358660	0.115438	-3.107	0.00189
Age20	-0.446384	0.107402	-4.156	3.24e-05
Age21	-0.346106	0.114655	-3.019	0.00254
Age22	-0.679372	0.122341	-5.553	2.81e-08
Age23	-0.537443	0.113656	-4.729	2.26e-06
Age24	-0.594062	0.106577	-5.574	2.49e-08
Age25	-0.908516	0.109876	-8.269	$< 2 \times 10^{-16}$
Age26	-1.114204	0.120463	-9.249	$< 2 \times 10^{-16}$
Age27	-0.959696	0.115583	-8.303	$< 2 \times 10^{-16}$
Age28	-0.954083	0.106344	-8.972	$< 2 \times 10^{-16}$
Age29	-0.921755	0.107920	-8.541	$< 2 \times 10^{-16}$
Age30	-1.755344	0.123194	-14.249	$< 2 \times 10^{-16}$
Academic.Pressure	0.833197	0.016985	49.055	$< 2 \times 10^{-16}$
CGPA_levelB	0.097619	0.058575	1.667	0.09560
CGPA_levelA	0.249139	0.056528	4.407	1.05e-05
Study.Satisfaction2	-0.382682	0.068009	-5.627	1.83e-08
Study.Satisfaction3	-0.437248	0.067670	-6.461	1.04e-10
Study.Satisfaction4	-0.771979	0.065961	-11.703	$< 2 \times 10^{-16}$
Study.Satisfaction5	-0.988634	0.071024	-13.920	$< 2 \times 10^{-16}$
Have.you.ever.had.suicidal.thoughts..1	2.484906	0.044149	56.285	$< 2 \times 10^{-16}$
Work.Study.Hours	0.118958	0.005678	20.951	$< 2 \times 10^{-16}$
Financial.Stress	0.560143	0.015315	36.575	$< 2 \times 10^{-16}$
Family.History.of.Mental.Illness1	0.264168	0.041616	6.348	2.19e-10
Sleep.Duration2	-0.435505	0.058578	-7.435	1.05e-13
Sleep.Duration3	-0.386962	0.056466	-6.853	7.23e-12
Sleep.Duration4	-0.665860	0.059081	-11.270	$< 2 \times 10^{-16}$

Table 10: Coefficients from the GLMM for Depression with Logit Link Function.



	id	Gender	State	City	Degree	New_Degree	Age	Academic.Pressure	CGPA
	<int>	<fctr>	<fctr>	<chr>	<chr>	<fctr>	<fctr>	<dbl>	<dbl>
31	263	0	Jammu & Kashmir	Srinagar	BCA	1	28	5	5.88
79	523	0	West Bengal	Kolkata	BHM	1	27	4	6.75
102	676	1	Andhra Pradesh	Visakhapatnam	Class 12	3	19	3	5.68
104	687	1	Maharashtra	Pune	BCA	1	22	1	6.17
139	843	1	Madhya Pradesh	Bhopal	B.Ed	1	23	5	5.74
163	1022	1	Gujarat	Ahmedabad	Class 12	3	18	3	6.16
179	1126	1	Uttar Pradesh	Ghaziabad	M.Com	2	29	2	5.88
244	1499	1	Maharashtra	Pune	Class 12	3	18	5	8.70
288	1809	0	Maharashtra	Thane	BA	1	22	3	6.03
292	1843	1	Madhya Pradesh	Indore	Class 12	3	20	4	7.10

1-10 of 643 rows | 1-10 of 19 columns

Figure 16: Outliers Extracted after Model Plotting