Classifying Depression Among University Students with Logistic Regression

PSTAT 100: Data Science Concepts and Analysis

Group Members

- Valerie De La Fuente (valeriedelafuente)
- Matthew Arteaga (matthewarteaga)
- Phuc Lu (pdlu)
- William Nelson (williamnelson)
- Hayden Galletta (haydengalletta)

1 Abstract

2 Introduction

For this project, the Student Depression Dataset by Israel Campero Jurado was used for analysis. This data set was published to OpenML on March 12, 2025 and was retrieved on April 22, 2025. Link

This data contains anonymized information that is useful for studying depression levels among students. This data set contains features such as the students' information (age and gender), their academic performance (grades, and attendance), their lifestyles (sleep patterns, exercise, and social activities), mental health history, and how they would rate their depression on a standardized scale. The raw data file is in the CSV format. The data in this data are structured, where each row represents an individual student and the columns represent a specific variable.

Previously, those who studied this data set tended to be psychology researchers, data science, and educators. Their aims were to identify factors that contribute to student depression and to design early intervention strategies. This research is also interested in looking at student depression and potential contributing factors. Hence, for this project, we're interested in answering the following questions:

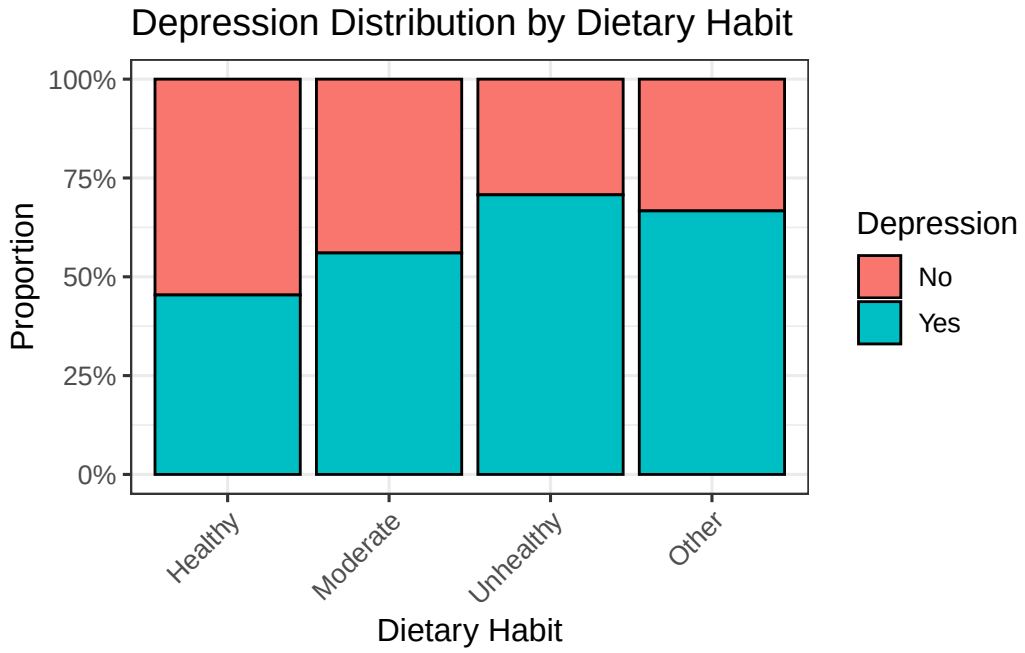
1. Do certain dietary habits coincide with an increased rate of depression among students?
2. Is there a correlation between the amount of sleep a student gets and the proportion of them that are depressed?
3. Does the presence (and magnitude) of certain stressors have an impact on the rate at which students are depressed?

We propose the following hypotheses:

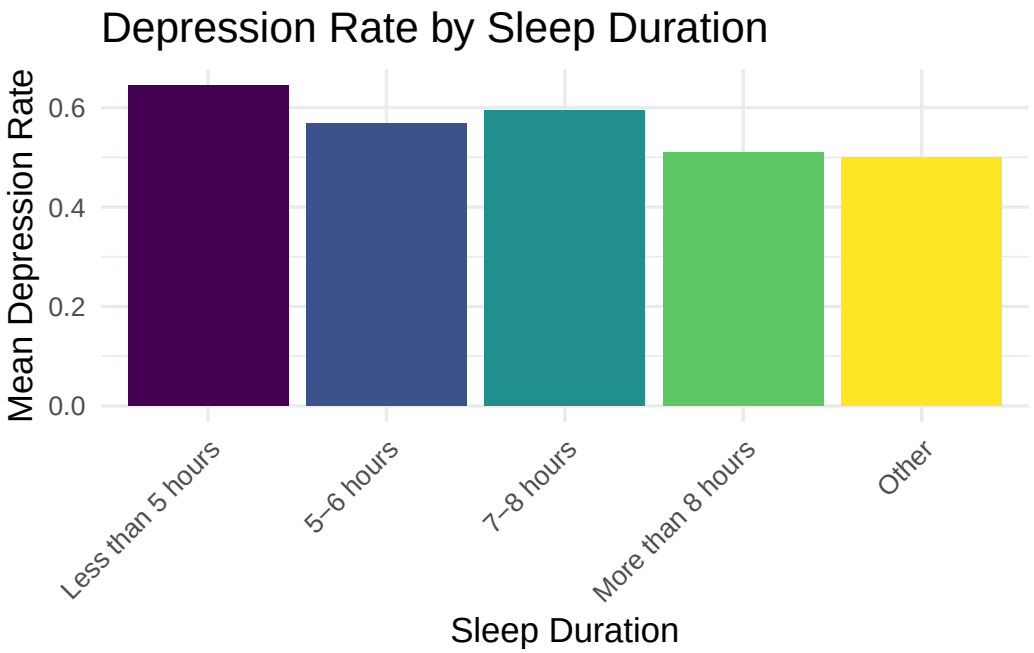
1. Students with moderate to healthy dietary habits will have lower rates of depression compared to students with unhealthy dietary habits.
2. Students who average more sleep per night will have lower rates of depression compared to students who average less.
3. Students with the highest collective reported stressors (academic pressure, work pressure and financial stress) will have higher rates of depression compared to students with lower collective reported stressors.

3 Exploratory Data Analysis

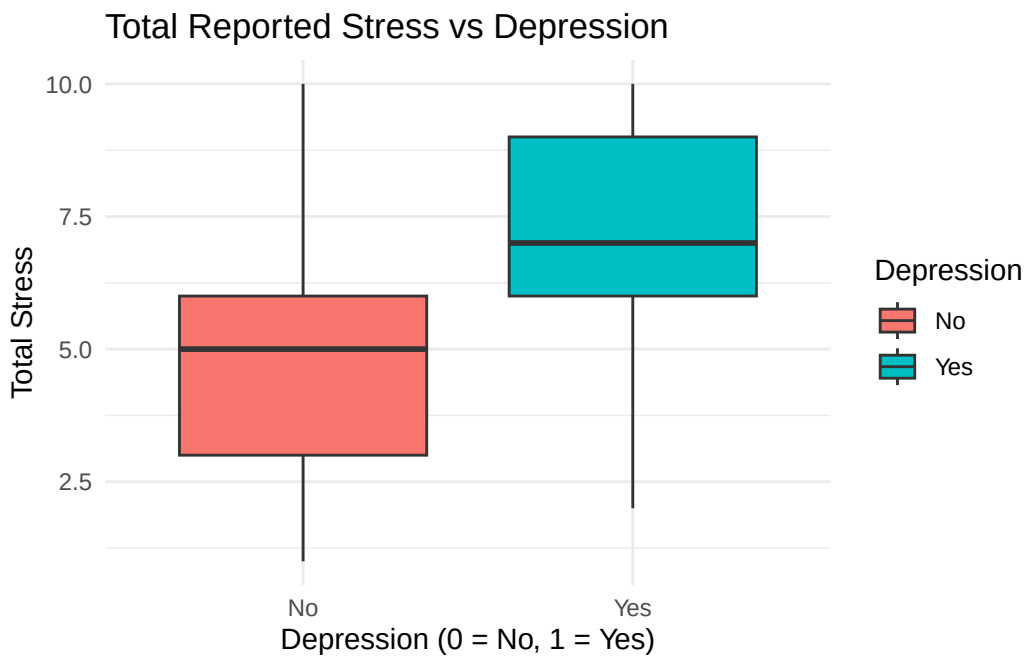
This first graph is a bar plot that helps us visualize hypothesis 1 by visualizing the correlation between a healthy diet and depression. Notably, around 45% of students with healthy dietary habits have depression, around 55% of students with moderate dietary habits have depression, and 75% of students with unhealthy dietary habits have depression. The results of this bar plot indicate that depression rates increase and student dietary habits worsen.



The second graph is a bar plot that helps to answer hypothesis 2 by visualizing the correlation between sleep patterns and depression. The results of this bar plot seem to indicate that people getting less than 5 hours of sleep have a significantly higher rate of depression and people who get more than 8 hours of sleep have a significantly lower rate of depression. This indicates that lower rates of depression are linked to a greater daily sleep duration.



This final graph is a boxplot that helps visualize hypothesis 3 and shows the correlation between multiple stressor factors and depression. The results seem to indicate a correlation with the increased stress levels and depression.



4 Data Processing

There are 27901 observations and 18 variables in this dataset. The column names in the original dataset were changed to the following: `id`, `gender`, `age`, `city`, `profession`, `academic_pressure`, `work_pressure`, `cum_gpa`, `study_satisfaction`, `job_satisfaction`, `sleep_duration`, `dietary_habits`, `degree`, `suicidal_thoughts`, `work_study_hours`, `financial_stress`, `fam_mental_illness`, and `depression`.

The following variables were mutated:

- `financial_stress`: variable type was converted to numeric and invalid entries were properly handled.
- `gender`: variable type was converted to a factor.
- `city`: typos and invalid entries properly handled.
- `profession`: invalid entries were properly handled.
- `sleep_duration`: invalid entries were properly handled and observations in the "Other" category were removed.
- `dietary_habits`: typos were properly handled.
- `suicidal_thoughts`: variable type was converted to a factor.
- `fam_mental_illness`: variable type was converted to a factor.
- `depression`: variable type was converted to a factor.

There are 3 missing observations present in the dataset, coming from the `financial_stress` variable. This accounts for less than 0.1% of the dataset, so we can easily remove them. With the data processed and tidied, we can move onto modeling.

5 Modeling Process

With 3 hypotheses to test, the modeling process will be outlined and broken into 3 parts.

5.1 Hypothesis 1: Students with moderate to healthy dietary habits will have lower rates of depression compared to students with unhealthy dietary habits.

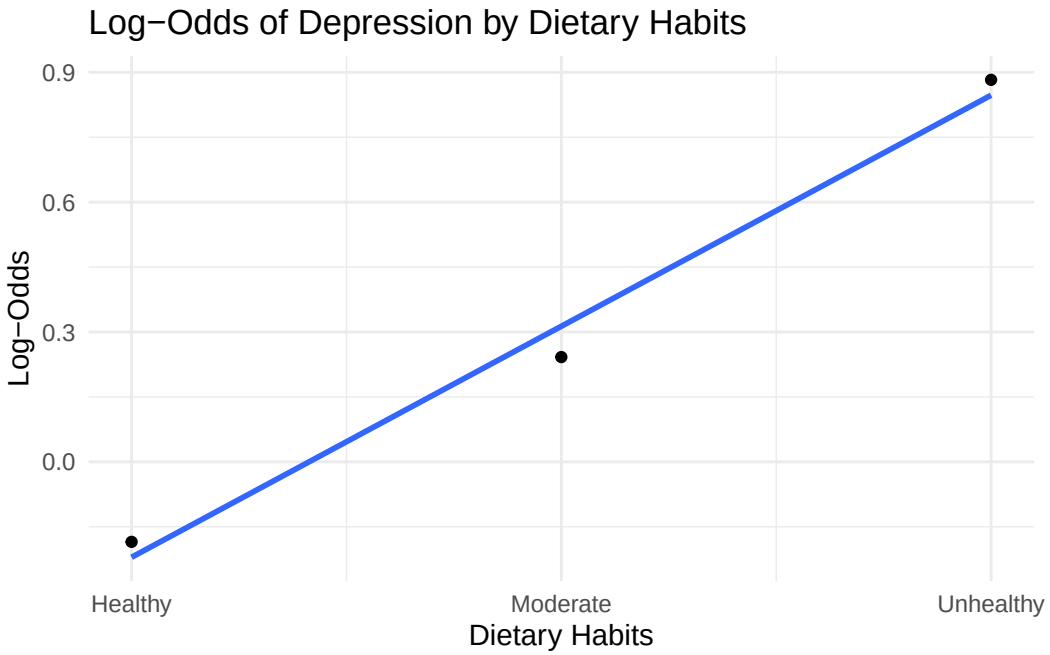
Our first hypothesis we want to test is that students with moderate to healthy dietary habits will have lower rates of depression than students with unhealthy dietary habits. We will fit a basic logistic regression model to predict the binary outcome variable, `depression`, using our categorical variable `dietary_habits` as a predictor.

Before fitting our model, we should first fully understand the predictor we are working with as well as check some assumptions that need to be met for our model to function properly. Our predictor `dietary_habits` is distributed as follows:

Habits	Count
Healthy	7649
Moderate	9921
Unhealthy	10316
Other	12

We have a slightly skewed distribution of responses between "healthy", "moderate", and "unhealthy" dietary habits, as well as 12 observations that responded "other". Because these 12 responses are a very small fraction of the overall data, we can remove these to simplify our model and our interpretations of it.

Logistic regression models rely on several key assumptions to perform correctly. Most of these assumptions have already been met or are reasonably assumed to be satisfied, except for the assumption that the predictor variable should have a linear relationship with the log-odds of the outcome variable. To assess this, we can plot the log-odds of the outcome against the levels of the predictor to evaluate whether the linearity assumption holds.



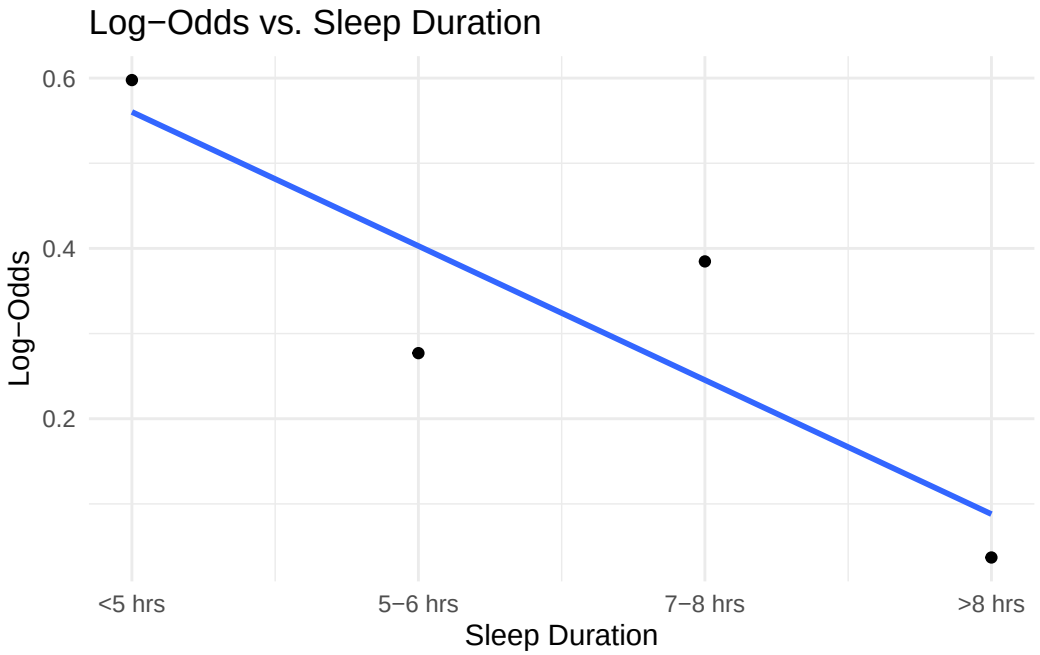
There appears to be a strong linear relationship between the log-odds of depression by each level of `dietary_habits`, so we can conclude that our data meets the necessary assumption and move on to fitting our model.

5.2 Hypothesis 2: Students who average more sleep per night will have lower rates of depression compared to students who average less.

Based on the hypothesis and the characteristics of the data set, our analytical approach of choice for investigation is classification; where we will construct a logistic regression model in an attempt to predict whether or not a student

reports experiencing depression based on how many hours of sleep they average per night. This is the best method of choice because our outcome variable, whether or not the student is depressed, is a binary value and our predictor, the amount of sleep averaged per night, is categorical with an ordinal nature. Additionally, the method quantifies associations and predicts probabilities, and can be extended to control for other factors.

Assumptions Required for Logistic Regression: In order to use logistic regression to investigate our hypothesis, there are a few assumptions of the data that must be met in order for the model to be valid. That is (1), the outcome variable is binary (condition is met), (2), that the observations are independent of one another (condition is assumed based on how data was collected), (3), that the log-odds of the outcome is a linear function of the predictor variable, (4), that there is no multicollinearity (not of concern; only one variable involved in model), and (5), that there at least 10 events per predictor level (condition is met). In our case the only assumption that needs to be checked is the linearity of the log-odds.



The graph of log-odds vs. sleep duration shows us a somewhat clear linear relationship between sleep duration and the log-odds. To investigate further, we will use a Box-Tidwell Test and look at the p-value corresponding to `sleep_log` (the log of the `sleep_duration` variable)

Table 2: GLM Coefficient Estimates

	z value	Pr(>|z|)
(Intercept)	6.051	0.000
`Sleep Duration`	-1.437	0.151
sleep_log	-0.065	0.948

Based on the p-value of 0.948 corresponding to the `sleep_log` variable, at significance level $\alpha = 0.05$, we fail to reject the null hypothesis that the log odds is a linear function of the `sleep_duration` predictor variable, thus the (3) assumption is met and we can proceed to constructing our model.

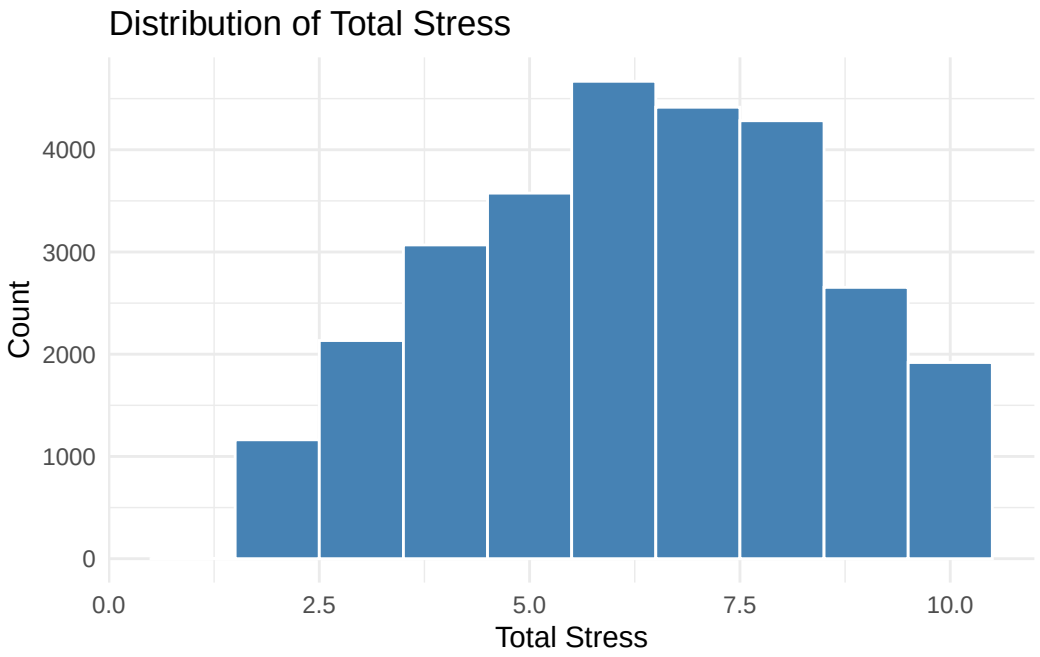
We will use the `glm` function to generate the following logistic regression model for predicting the proportion of students reporting depression at the varying sleep ranges (and how that may be extrapolated to see how depression rates scale with sleep duration):

$$\hat{Depression Proportion} = 0.72583 - 0.15739 Sleep Duration$$

5.3 Hypothesis 3: Students with the highest collective reported stressors (academic pressure, work pressure and financial stress) will have higher rates of depression compared to students with lower collective reported stressors.

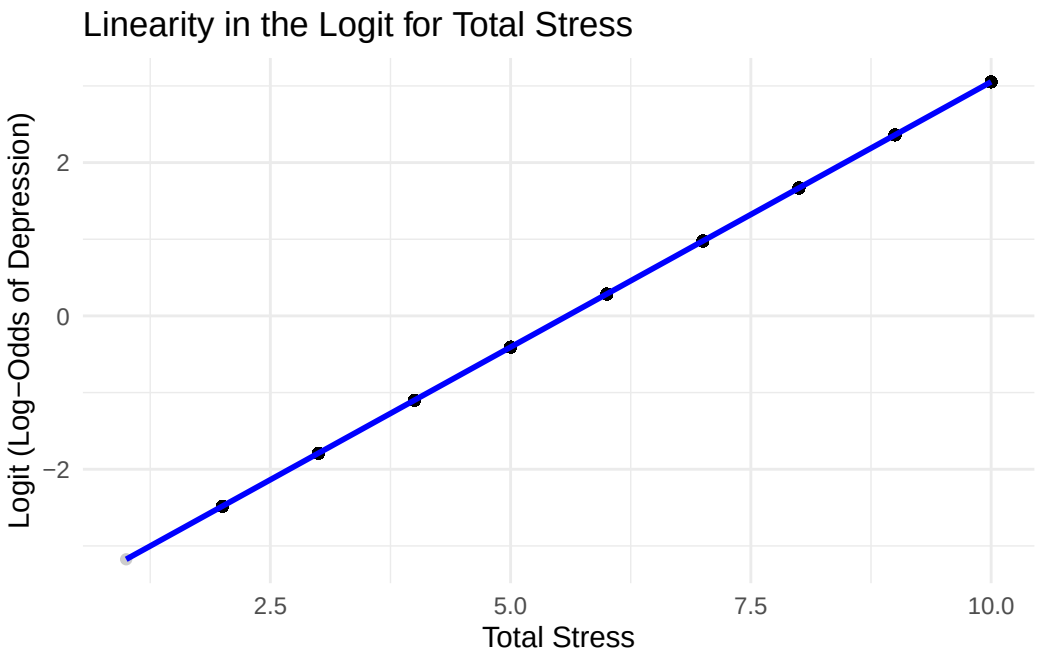
Our final hypothesis we want to test is that students with multiple combined stressors will have a higher overall rate of depression than students without these stressors. We will fit a logistic regression model to predict the binary outcome variable, `depression`, using the numeric variable `total_stress` (combination of `academic_pressure` + `work_pressure` + `financial_stress`) as a predictor.

Before fitting our model, we should first fully understand the predictor we are working with as well as check some assumptions that need to be met for our model to function properly. Our predictor `total_stress` is distributed as follows:



Looking at our data we can tell that it has a bell curve distribution and overall has a good representation of many different `total_stress` values. As a result the data does not need any cleaning and is ready to be used in our logistic regression model.

As mentioned in the previous hypothesis, logistic regression models have a few assumptions that we assume to have already met. The only one that we need to verify is the one that says the predictor variable should be linear in the log-odds of the outcome variable. We can check this by plotting the log-odds against the different amounts of `total_stress`:



There appears to be a strong linear relationship between the log-odds of `depression` by `total_stress`, so we can conclude that our data meets the necessary assumption and move on to fitting our model.

6 Results and Interpretation

6.1 Hypothesis 1: Students with moderate to healthy dietary habits will have lower rates of depression compared to students with unhealthy dietary habits.

We fit a logistic regression model with the logit link function that will predict the odds of having depression by each level of `dietary_habits`. The results of the model are summarized in the following table:

	Estimate	Std. Error	z value	Pr(>|z|)
dietary_habits	0.2126795	0.005557574	38.26841	0

We see statistically significant results from our basic logistic regression model with just dietary habits as a predictor. Every level of dietary habits has statistically significant effects on the presence of depression. Our model predicts that "healthy" dietary habits decrease the probability of depression by about 18.5%, while "moderate" and "unhealthy" dietary habits increase the probability of depression by about 24% and 88% respectively. This is a very strong result to start with, but we should check our model results in other ways as well.

We will first check to make sure the `dietary_habits` categorical predictor is significant in predicting the presence of depression overall rather than just by the level of the variable. We will use the `anova()` function to perform a chi-squared likelihood ratio test with a null hypothesis that the null model without `dietary_habits` is sufficient in predicting depression, and an alternative hypothesis that `dietary_habits` is significant to the model in terms of lowering the model's residual deviance. The results of the ANOVA are the following:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	27886	38658.20	NA
dietary_habits	1	1523.597	27885	37134.61	0

This table shows us that when added to the null model, the `dietary_habits` predictor reduced the model's residual deviance so greatly that the p-value for our likelihood ratio test is too small for R to display it. The true p-value is

less than 2.2 * 10^-16, which is extremely tiny and reasonably rounded to 0. With this similarly significant result to the model results earlier, we can conclude that the Dietary Habits predictor as a whole is significant in predicting the presence of depression.

We can now check how accurate our model is at predicting the presence of depression using a confusion matrix. The confusion matrix will visualize the accuracy of our model in terms of true positive and negative rates in the diagonal cells, false positive rate in the bottom left cell (row 2, column 1), and false negative rate in the top right cell (row 1, col 2).

	0	1
0	0.000	0.000
1	0.415	0.585

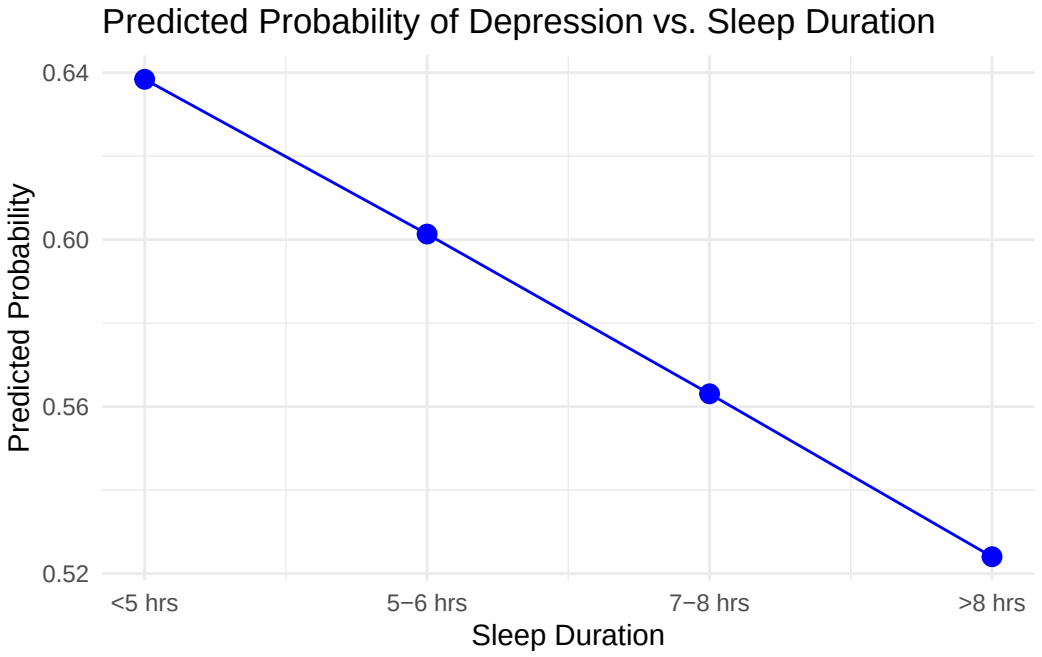
Here we see our model has an accuracy of 61%, a false positive rate of 26.5%, and a false negative rate of 12.5%. This means our model with just `dietary_habits` as a predictor is slightly better at predicting the presence of depression than a random guess.

Despite the fact that this model cannot predict the presence of depression with an impressive accuracy, the results of our model fit and surrounding hypothesis tests lead us to fail to reject our hypothesis that students with moderate to healthy dietary habits will have lower rates of depression than students with unhealthy dietary habits.

6.2 Hypothesis 2: Students who average more sleep per night will have lower rates of depression compared to students who average less.

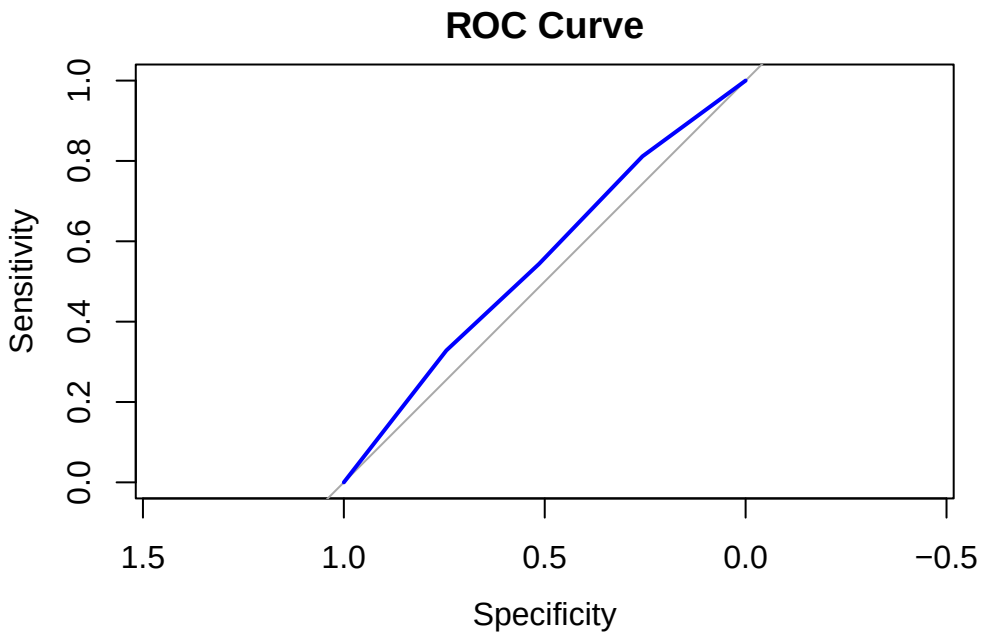
Table 6: Logistic Model Table

term	estimate	std.error	statistic	p.value
(Intercept)	0.726	0.029	24.969	1.32e-137
`Sleep Duration`	-0.157	0.011	-14.510	1.04e-47



The model has an intercept coefficient of 0.72583, representing the average depression rate for students falling under the `0` Sleep Range (irrelevant as `1` is the reference group; representing less than 5 hours of sleep), and a `sleep_duration` coefficient of -0.15739, representing the average change in depression probability when going from one sleep range to the next (in order). The p-values for both coefficients are <0.05, indicating statistical significance of the model.

To analyze the performance of the model we will investigate the ROC curve and area under the curve generated by the model:



The ROC curve generated by the model is slightly above the diagonal (increasing and concave down), but not by much. Additionally, the AUC generated from the graph is 0.5494. This means that the prediction made by the model is relatively random, skewing slightly towards being a good model (correctly predicting the depression proportion based on `sleep_duration`).

6.3 Hypothesis 3: Students with the highest collective reported stressors (academic pressure, work pressure and financial stress) will have higher rates of depression compared to students with lower collective reported stressors.

We fit a logistic regression model that will predict the odds of having depression by each level of `total_stress`. The results of the model are summarized in the following table:

Table 7: Logistic Regression: Predicting Depression from Total Stress

	Term	Estimate	Std..Error	z.value	Pr…z..	Odds.Ratio
(Intercept)	(Intercept)	-3.866	0.054	-71.64	0.00e+00	0.021
total_stress	total_stress	0.692	0.009	79.50	0.00e+00	1.997

We see statistically significant results from our basic logistic regression model using `total_stress` as a predictor. Our model predicts that for every one point increase in the presence of `total_stress` will double the likelihood of depression. This is a very strong result to start with, but we should check our model results in other ways as well.

We will first check to make sure the `total_stress` numeric predictor is significant in predicting the presence of depression overall. We will use the `anova()` to perform a chi-squared likelihood ratio test with a null hypothesis that the null model without `total_stress` is sufficient in predicting depression, and an alternative hypothesis that `total_stress` is significant to the model in terms of lowering the model's residual deviance. The results of the ANOVA are the following:

Table 8: Chi-Squared Likelihood Ratio Test: Comparing Null vs. Total Stressors Model

Model	Df	Deviance	Residual_Df	Residual_Dev	P_value
1	NA	NA	2.7885e+04	3.783894e+04	NA
2	1e+00	9.61814e+03	2.7884e+04	2.822080e+04	0e+00

This table shows us that when added to the null model, the `total_stress` predictor reduced the model's residual deviance so greatly that the p-value for our likelihood ratio test is too small for R to display it. The true p-value is less than 2.2e-16, which is extremely tiny and reasonably rounded to 0. With this similarly significant result to the model results earlier, we can conclude that the `total_stress` predictor as a whole is significant in predicting the presence of depression. Thus we reject the null hypothesis that the simpler model without `total_stress` is sufficient in predicting depression. The extremely low p-value indicates that the model including `total_stress` provides a significantly better fit to the data. Thus, we conclude that `total_stress` is a significant predictor of depression.

Lastly we will check how accurate our model is at predicting the presence of depression using a confusion matrix. The confusion matrix will visualize the accuracy of our model in terms of true positive and negative rates in the diagonal cells, false positive rate in the bottom left cell (row 2, column 1), and false negative rate in the top right cell (row 1, col 2).

Table 9: Confusion Matrix (Predicted vs Actual)

Prediction	Actual: No	Actual: Yes
No	7356	2587
Yes	4203	13740

Table 10: Model Performance Metrics from Confusion Matrix

Metric	Value
Accuracy	0.7565
95% CI	(0.7514, 0.7615)
No Information Rate	0.5855
P-Value [Acc > NIR]	< 0.001
Kappa	0.4879
McNemar's Test P-Value	< 0.001

Here we see our model has an accuracy of 75.65%. This means our model with `total_stress` as a predictor is significantly better at predicting the presence of depression than a random guess.

Seeing that this model can predict the presence of depression with an impressive accuracy, the results of our model fit and surrounding hypothesis tests lead us to reject the null hypothesis that the simpler model without `total_stress` is sufficient in predicting depression. Thus concluding that the `total_stress` is a significant predictor of depression.

7 Conclusion and Recommendations