Step 3 Hypothesis 2

Matthew Arteaga

Hypothesis 2

- **Research Question**: Is there a correlation between the amount of sleep a student gets and the proportion of them that are depressed?
- **Hypothesis**: Students who average more sleep per night will have lower rates of depression compared to students who average less.

1.1 Data Analysis

Based on the research question, 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.

Data Processing

- **Data Cleaning**: In order to properly clean our data to test this hypothesis we must check to see if there are any missing values for both the outcome variable (`Depression`) and our chosen predictor variable (`Sleep Duration`):

Table 1: NA Values

Variable	Missing_Count
Depression	0
Sleep Duration	0

- After aggregating the missing values data for both the `Depression` and `Sleep Duration` variables, we found that there are no `NA` values assigned to any observations for the respective variables. However this does not mean that there are no missing values present in the data. In order to investigate further we need to look at a frequency table of both variables:

Table 2: Sleep Frequency

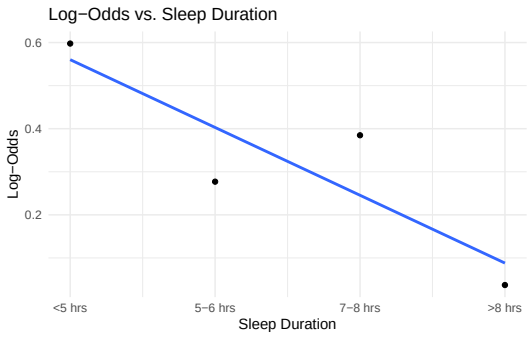
Variable	Frequency
'5-6 hours'	6183
'7-8 hours'	7346
'Less than 5 hours'	8310
'More than 8 hours'	6044
Others	18

Table 3: Depression Frequency

Variable	Frequency
0	11565
1	16336

- As evident by the frequency tables, there are no missing values for the `Depression` variable in the data set, however there are 18 instances of `Others` being listed as values for the `Sleep Duration` variable, so we will treat those as instances with missing values and remove them from our data set.

- **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 4: 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.

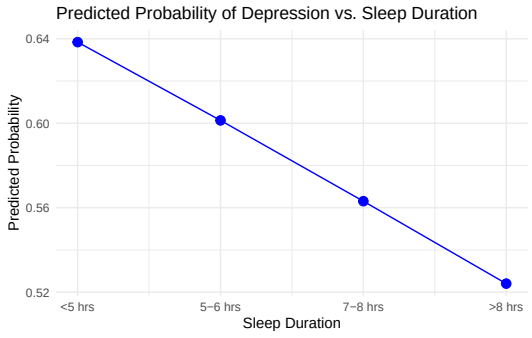
2.1 The Model

- **Results**: We 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):

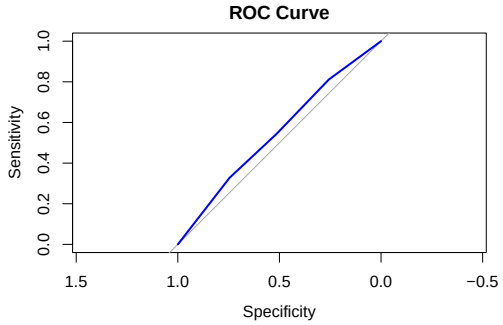
$$\hat{Depression Proportion} = 0.72583 - 0.15739 SleepDuration$$

Table 5: 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



- **Interpretation**: 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.
- **Analyzing Performance: ROC curve and AUC**: 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).