# Final Step 3

## Abstract

## Introduction

## Hypothesis 1: Will

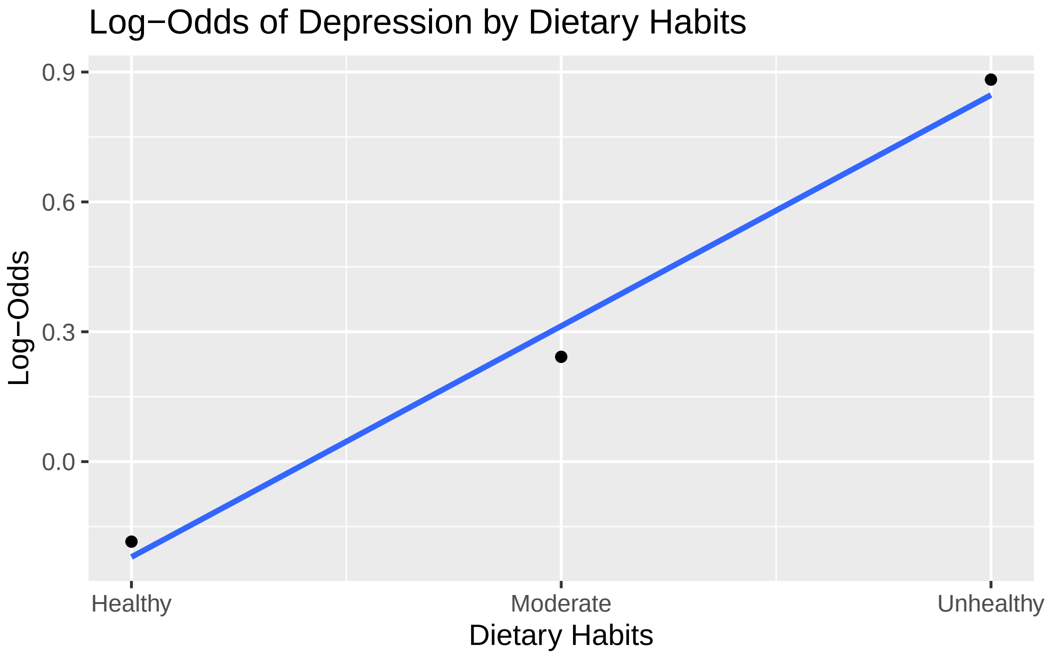
Our first hypothesis we want to test is that students with moderate to healthy dietary habits will have lower rates of depression than students with unhealthy dietary habits. We will fit a basic logistic regression model to predict the binary outcome variable, “depression”, using our categorical variable “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 have a few assumptions that must be met to perform properly. (Named elsewhere) We already have met, or assume we have met, most of the assumptions besides 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 levels of the predictor variable to assesst the linearity:



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.

We will 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 percent 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 |
|---|---|---|---|---|
| NULL | NA | NA | 27886 | 38658.20 |
| dietary_habits | 1 | 1523.597 | 27885 | 37134.61 |

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 0.61 or 61%, a false positive rate of 0.265 or 26.5%, and a false negative rate of 0.125 or 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.

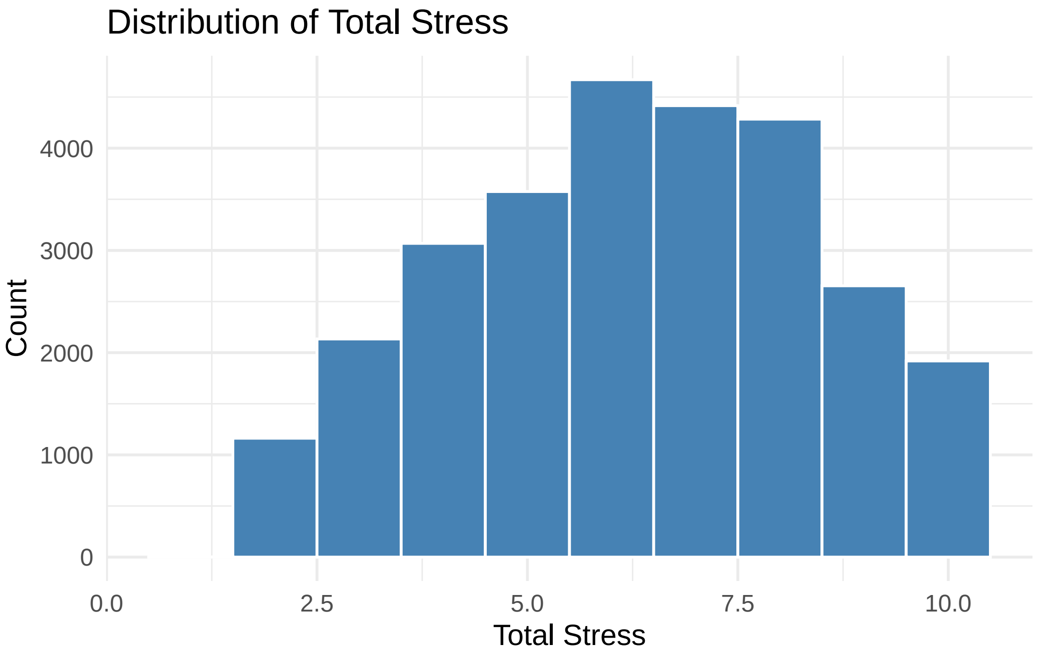
### Hypothesis 2: Matthew

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

### Hypothesis 3: Hayden

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 Stressors (combination of Academic Pressure + Work Pressure + Financial Stress) as a predictor.

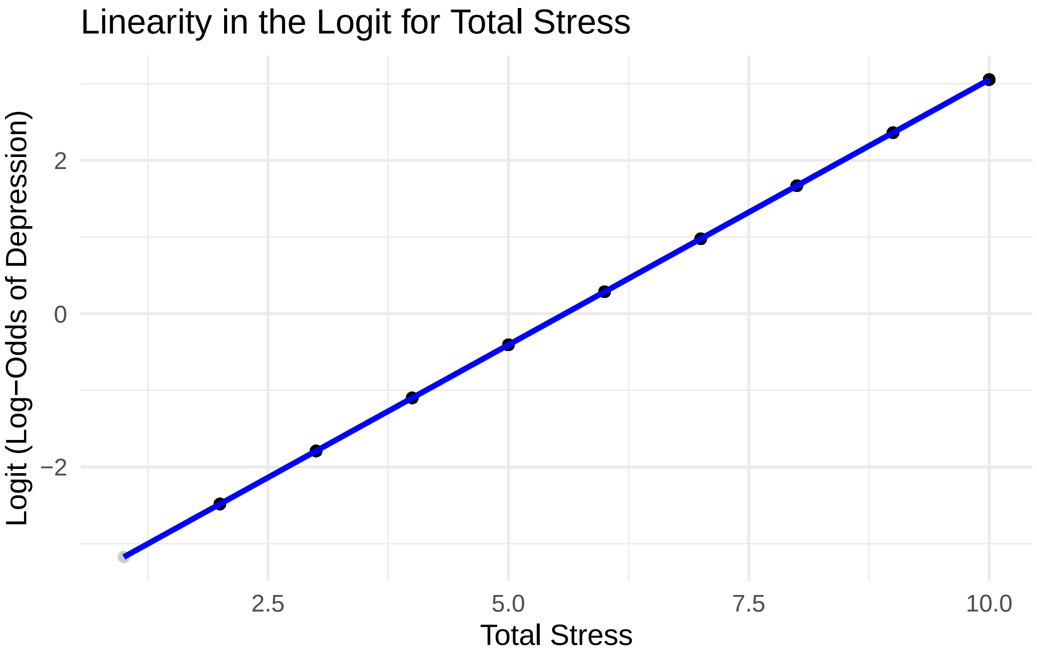
Before fitting our model, we should first fully understand the predictor we are working with aswell as check some assumptions that need to be met for our model to function properly. Our predictor "Total Stressor" is distributed as follows:



.

Looking at our data we can tell that it has a bell curve distribution and overall has a good representation of many different Total Stressor 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 Stressors:



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

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

Table 5: 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 Stressors as a predictor. Our model predicts that for every one point increase in the presence of Total Stressors 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 Stressors numeric predictor is significant in predicting the presence of depression overall. We will use the "anova" function to perform a Chi-Squared likelihood ratio test with a null hypothesis that the null model without Total Stressors is sufficient in predicting depression, and an alternative hypothesis that Total Stressors is significant to the model in terms of lowering the model's residual deviance. The results of the anova are the following:

Table 6: 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 Stressors 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 Stressors predictor as a whole is significant in predicting the presence of depression. Thus we reject the null hypothesis that the simpler model without Total Stressors 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 7: Confusion Matrix (Predicted vs Actual)

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

Table 8: 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 0.7565 or 75.65%. This means our model with Total Stressors 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 Stressors is sufficient in predicting depression. Thus concluding that the Total Stressors is a significant predictor of depression.

## Conclusion

## Recommendations