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.

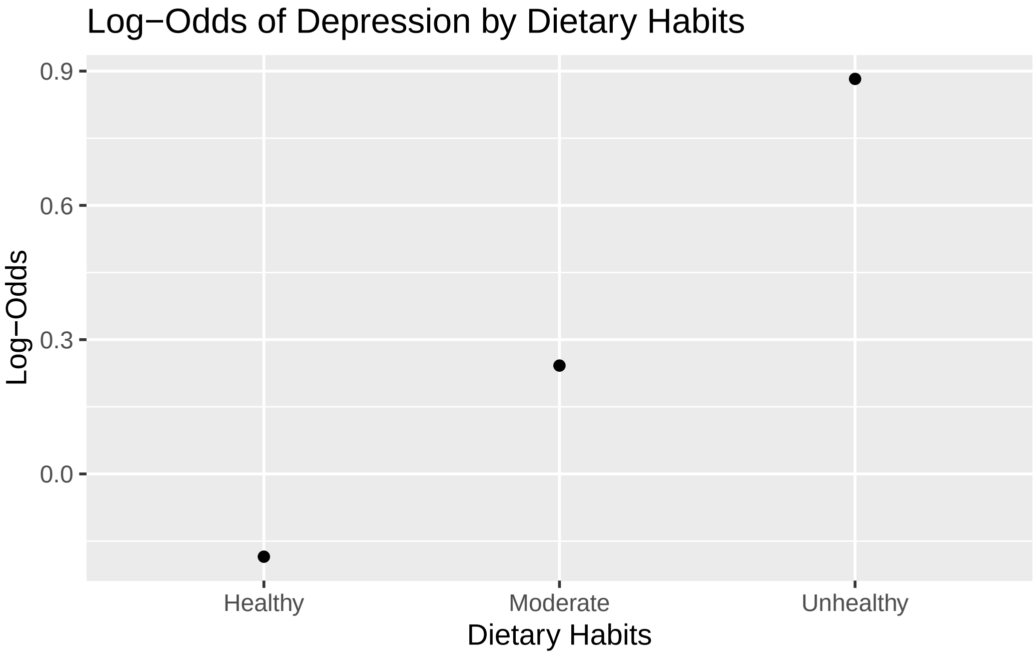
Logistic Regression models have a few assumptions that must be met to perform properly. (Named elsewhere) We already have met, or assume we have met, most of the assumptions besides the one that says the predictor variable should be linear in the log-odds of the outcome variable. We can check this by plotting the log-odds against the levels of the predictor variable to assesst the linearity:

```
depression_data <- depression_data %>% mutate(
  depression = case_when(
    depression == "Yes" ~ 1,
    depression == "No" ~ 0
  )
)
```

```
logOddsDF <- depression_data %>%
  group_by(dietary_habits) %>%
  summarize(
    odds = mean(depression),
    logOdds = log(odds/(1-odds)))

ggplot(logOddsDF, aes(x = dietary_habits, y = logOdds)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Log-Odds of Depression by Dietary Habits", y = "Log-Odds", x = "Dietary Habi
```

```
`geom_smooth()` using formula = 'y ~ x'
```



The results of the model are summarized in the following table:

	Estimate	Std. Error	z value	Pr(>|z|)
dietary_habitsHealthy	-0.1848625	0.02296573	-8.049494	8.313729e-16
dietary_habitsModerate	0.2420784	0.02022674	11.968239	5.212305e-33
dietary_habitsUnhealthy	0.8825377	0.02163972	40.783224	0.000000e+00

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	Pr(>Chi)
NULL	NA	NA	27886	38658.20	NA
dietary_habits	3	2038.137	27883	36620.07	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.

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

Students with the highest collective reported stressors (Academic Pressure + Work Pressure + Financial Stress) will have higher rates of depression compared to students with lower collective reported stressors.

Conclusion

Recommendations