Project 2

Ethan Lewis

10/31/2021

Executive Summary

Understanding exercise habits within a population is valuable information, not only to marketing companies working closely with fitness brands, but also to investigating underlying health trends. In the report below, we outline our process for constructing a predictive exercise model along with a through analysis of our findings and their limitations.

Section 1: Introduction

To predict (ir)regularly exercising individuals in the United States, we will explore a subset of the Behavioral Risk Factor Surveillance System (BRFSS): an annual national health survey designed to identify risk factors in the adult population and report emerging health trends. Our particular subset contains 20,000 individuals worth of information on 9 variables:

- General Health Self Evaluation (Excellent/Very Good/Good/Fair/Poor)
- Exercise Within Past Month (Y/N)
- Health Coverage (Y/N)
- Previously Smoked 100 Cigarettes (Y/N)
- Height (in.)
- Weight (lbs.)
- Desired Weight (lbs.)
- Age
- Gender (M/F)

Section 2: Exploratory Data Analysis

First, let's thoroughly explore our response variable, whether or not an individual exercised in the past month, using Table 1 below where 1 represents "Yes" and 0 represents "No".

Table 1: 'exerany' Distribution

Exercise	Count
0	5086
1	14914

Of our 20,000 surveyed individuals, about 25% did not exercise within the past month while the other 75% did. Although this percentage disparity is noticeable, we have a sufficient number of "No" responses to comfortably proceed with the remaining EDA and modeling processes.

Next, we need to explore the relationships between the log(odds) of whether or not an individual exercised in the past month and our explanatory variables. Let's start with the numeric explanatory variables (height, weight, wtdesire, age) using Figures 2.1-4 below, respectively.

Figure 2.1

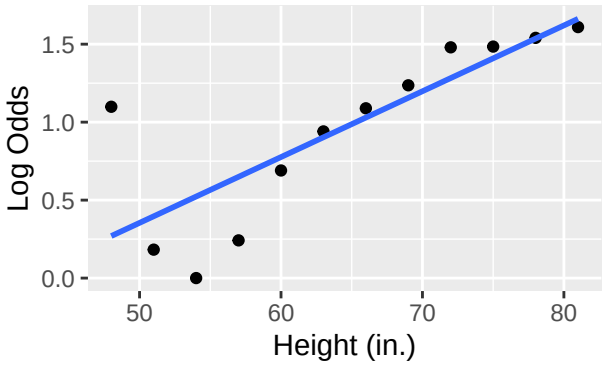


Figure 2.2

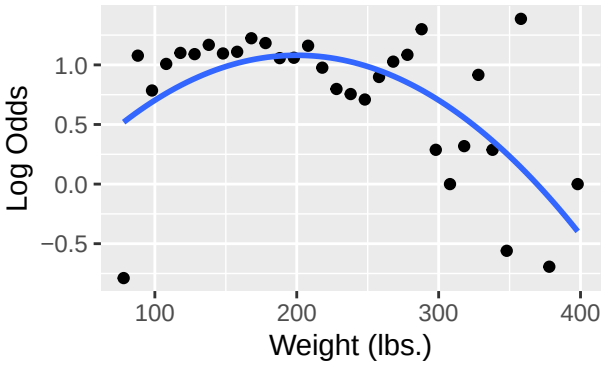


Figure 2.3

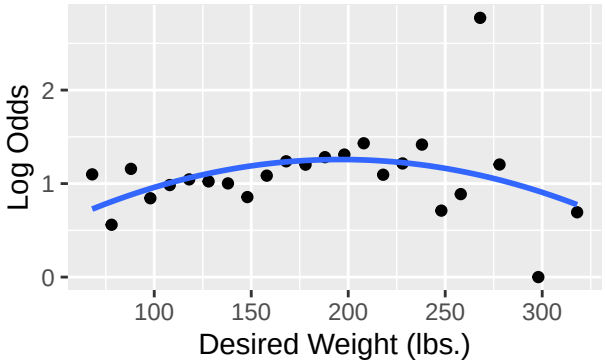
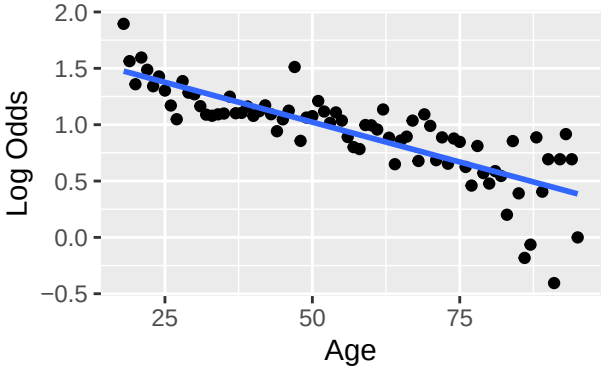


Figure 2.4



In Figure 2.1 we notice, for the most part, a positive linear trend between the log(odds) of exercising and height. There is certainly an outlier around 48 inches, but we chose to ignore this point when identifying and calculating the trend. Adult heights under 4 feet are uncommon and these individuals are likely exercising out of necessity. Without this point, a strong, positive linear trend remains in tact.

Figure 2.2 demonstrates an obvious negative 2nd order polynomial relationship between the log(odds) of exercise and weight. The trend begins to slightly unravel for heavier individuals as those with similar exercising log(odds) to lighter individuals are likely trying to lose weight. In fact, this phenomena seems to be evident in Figure 2.3: the bump in the trend line represents a common weight heavy individuals are looking to achieve.

Finally, Figure 2.4 blatantly depicts a negative linear relationship between the log(odds) of exercising and age: as we get older, we usually exercise less.

This numeric variable EDA gives us an indication height, weight, and age are predictors we need to consider, while wtdesire may not be particularly as important. Now let's examine the categorical explanatory variables using Figures 2.5-8, below.

Figure 2.5

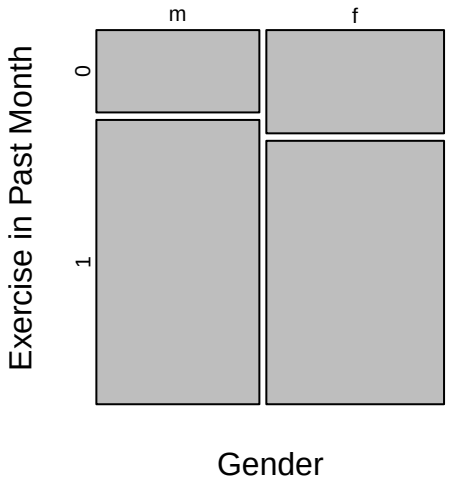


Figure 2.6

First, Figures 2.5 and 2.6 depict no real difference between the gender of an individual, as well as whether or not an individual previously smoked, as it relates to that individual exercising within the past month. Perhaps these variables should be excluded from the model.

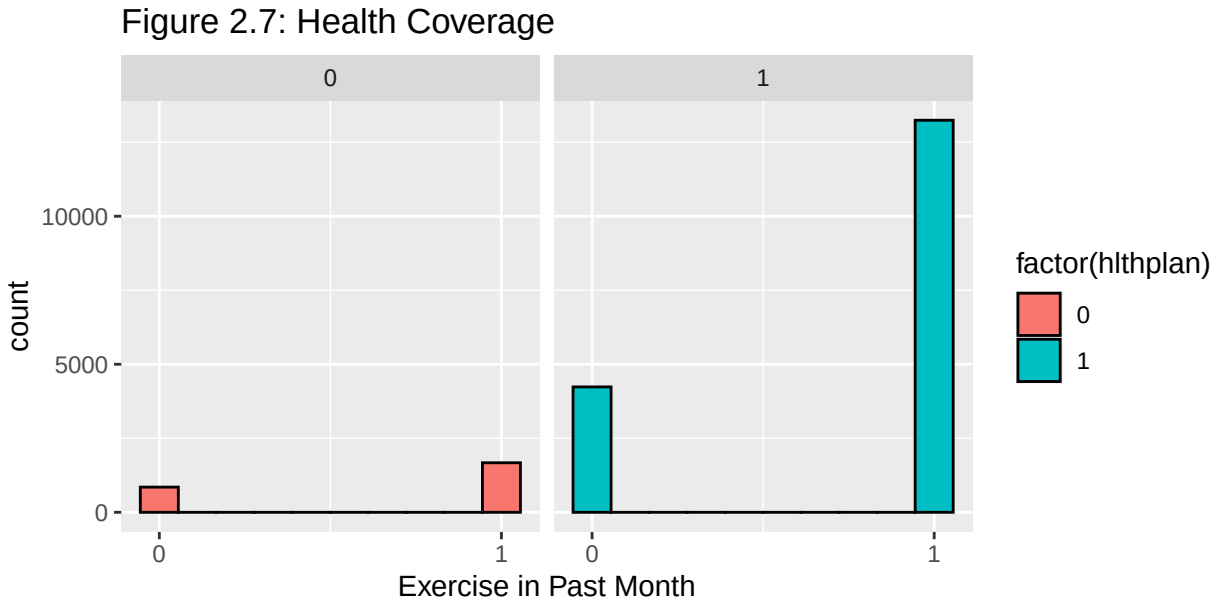


Figure 2.7, however, tells an opposite story as we notice an obvious difference between individuals with and without health coverage in regards to their exercise pattern. Those with some form of health coverage vastly outnumber those without when it comes to the odds of having exercised within the past month.

Figure 2.8

Finally, Figure 2.8 indicates a clear trend between exercise and an individual's health self-evaluation. In other words, as the positivity of an individual's self-evaluation decrease, so too does their odds of exercising. Notice the "very good" self-evaluation level has been absorbed by the "excellent" level since these were extremely similar in our initial mosaic plot.

Section 3: Modeling

Having completed a thorough EDA on all of our variables, we are ready to begin constructing some models. Specifically, we will explore two models: a full model and a reduced model comprised of the influential variables from EDA.

Since we are investigating a singular, binary variable, a Bernoulli distribution with parameter π_i is appropriate. The population form of our full model, Model 1, is below.

$Y_i \sim \text{Bernoulli}(\pi_i)$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0+\beta_1(Height_i)+\beta_2(Weight_i)+\beta_3(Weight_i^2)+\beta_4(DesiredWeight_i)+\beta_5(DesiredWeight_i^2)+\beta_6(Age_i)+$$

$$\beta_7(GenderF_i)+\beta_8(HealthCoverage_i)+\beta_9(Smoke100_i)+\beta_{10}(GoodHealth_i)+\beta_{11}(FairHealth_i)+\beta_{12}(PoorHealth_i)$$

Additionally, a collection of Model 1's fitted metrics are listed below.

- Residual Deviance = 21315
- AIC = 21341

- BIC = 21443.71

Next, let's create a reduced model called Model 2. Based on our EDA, Model 2 will exclude wtdesire, gender, and smoke100 resulting in the population form found below

$Y_i \sim$ Bernoulli(π_i)

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(Height_i) + \beta_2(Weight_i) + \beta_3(Weight_i^2) + \beta_4(Age_i) +$$
$$\beta_5(HealthCoverage_i) + \beta_6(GoodHealth_i) + \beta_7(FairHealth_i) + \beta_8(PoorHealth_i)$$

Additionally, a collection of Model 2's fitted metrics are listed below.

- Residual Deviance = 21322
- AIC = 21340
- BIC = 21411.14

Model 1 and 2 have extremely similar metrics: Model 1 wins in Residual Deviance, Model 2 in BIC, and essentially a tie in AIC. This indicates we need to perform a Nested Likelihood Ratio Test to better compare the two models. The steps for this hypothesis test at an α of 0.05 are listed below.

(1) H_O : Model 1, the full model, is not a better fit
H_A : Model 1, the full model, is a better fit

(2) $\widehat{D_{Model1}} = 21315$ where $p_{Model1} = 13$
$\widehat{D_{Model2}} = 21322$ where $p_{Model2} = 9$

(3) G = 21322 - 21315 = 7

(4) G ~ $\chi^2(4)$

(5) We have a p-value of 0.1358882 meaning if H_O were true, the probability of seeing a drop in deviance value of 7 is about 13.589%

(6) Based on this large p-value, we do not have convincing evidence Model 1 isn't a better fit than Model 2

Despite the conclusions of our hypothesis test, we would like to proceed with the reduced model, Model 2. In reality, Models 1 and 2 are extremely similar and would likely yield comparable results; our decision was made on the basis of the BIC (and simultaneously the number of parameters). The models share an identical AIC value, but after evaluating a similar metric (BIC) we notice Model 2 is a slight improvement. Much like the BIC metric, we ultimately preferred the model involving fewer parameters and, once fit, Model 2 took the equation below.

$$log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -3.146 + 0.07545(Height_i) - 0.001794(Weight_i) - 0.000007715(Weight_i^2) - 0.006703(Age_i) +$$
$$0.4392(HealthCoverage_i) - 0.5714(GoodHealth_i) - 0.9747(FairHealth_i) - 1.534(PoorHealth_i)$$

We can see all of the relationships identified in EDA at work in Model 2's equation. The negative slopes associated with weight, age, and the genhlth levels indicate a decrease in the log(odds) of exercising as these parameter values increase and vice versa for the positive slopes of height and hlthplan.

Finally, let's generalize our model a bit more by computing 95% confidence intervals for all parameters involved. These intervals are listed below.

- β_0: (-3.80561, -2.48915)
- $\beta_1(Height_i)$: (0.06542, 0.085501)
- $\beta_2(Weight_i)$: (-0.006364, 0.002833)
- $\beta_3(Weight_i^2)$: (-0.00001932, 0.000003697)
- $\beta_4(Age_i)$: (-0.008716, -0.0046884)
- $\beta_5(HealthCoverage_i)$: (0.34246, 0.535313)
- $\beta_6(GoodHealth_i)$: (-0.64672, -0.49592)
- $\beta_7(FairHealth_i)$: (-1.0791, -0.87012)
- $\beta_8(PoorHealth_i)$:(-1.69902, -1.36959)

Section 4: Prediction

Let's quickly test our model on an individual who is in good health, has health coverage, does not smoke, is 72 inches tall, weights 190 pounds, desires to weigh 180 pounds, is 26, and identifies as biologically male. To do so, we have plugged the appropriate values into the equation of Model 2 and solved for $\hat{\pi_1}$.

$$\hat{\pi_1} = \frac{e^{-3.146+0.07545(72)-0.001794(190)--0.000007715(36100)-0.006703(26)+0.4392-0.5714}}{1+e^{-3.146+0.07545(72)-0.001794(190)--0.000007715(36100)-0.006703(26)+0.4392-0.5714}}$$

$$\hat{\pi_1} = 0.87186$$

This means our model is predicting an 87.186% chance this individual exercises regularly.

Extending this evaluation of prediction accuracy further, we can use our model and data frame to make 20,000 individual predictions and compare these predictions to the true value of the respective individuals 'exerany' variable. Importantly, we must convert any positive prediction value to a 1 and any negative prediction value to a 0. This is a result of Model 2 not predicting the exact outcome of whether or not an individual exercised within the past month, but rather the log(odds)/probability an individual exercised within the past month. The classification error rates of our predictions are listed below.

$CER_{total} = \frac{4553+425}{20000} = 24.89\%$

$CER_{NoExercise} = \frac{4553}{5086} = 89.52\%$

$CER_{Exercise} = \frac{425}{14914} = 2.85\%$

Section 5: Conclusion

Having completed a thorough analysis of the provided data, we were able to produce a concrete model in which we are fairly confident. The model is highlighted by it's slope values corroborated in EDA as well as it's low Total CER and False Negative rates (24.89% and 2.85%, respectively). Unfortunately, however, a major flaw in our model is it's extremely high False Positive Rate (89.52%). This is somewhat difficult to explain, but it is clear our model is much better at predicting those who exercise than those who don't (perhaps we are missing some key explanatory variables).

Future work on this model would mainly involve addressing the high False Positive Rate, however, we would also explore the '$weight^2$' parameter as it seems to be negligible. Overall, our current model should do a decent job of predicting individuals who exercise.