Project 1

Jessica Chen

Abstract

Many factors contributed to people’s health-related risk behaviors and chronic health conditions. And specifically, we are interested in studying the probability of people exercising regularly by asking whether the respondent exercised in the past month or not. The data we used comes from BRFSS, and we only decided to use the subset of this data. In the report, we performed analysis on every potential variable in the data, and fitted the model which can predict the probability of the respondent exercising regularly. We detailed the steps of the process, and compared our model with two other potential models of estimating the probability of exercising regularly. It turns out that the probability of exercising regularly has a negative relationship with respondent’s expected weight loss, age, and has a positive relationship with general health conditions, health plan, and height. Finally, we did a prediction test to test the accuracy of our model. Our model performs great in predicting the true positive, but needs improvements in predicting the true negative.

Section 1: Introduction

We are interested in two research questions. First, the relationship between how much weight someone wants to lose, and the probability that they exercise regularly, after accounting for their age, general health, and health coverage. And the second research interest is, how well we can predict whether a patient exercises regularly if using other variables in the data, such as gender and age.

The data we use comes from BRFSS - Behavioral Risk Factor Surveillance System, which is the nation’s premier system of health-related telephone surveys that collects data over 50 states on US resident’s health-related risk behaviors, chronic health conditions. Over 400,000 adults take questionnaires each year, specifically, they will be asked on their health status, alcohol consumption, immunization, frequency of eating fruit and vegetables etc.

We will use the subset of the data, focusing on a random sample of 20,000 people from the BRFSS survey, and analyze 9 of the total variables. We will first perform exploratory data analysis (EDA) on the chosen variables, and then we will fit the model, test the model accuracy and answer the two research interests stated above.

Section 2: Data

Data Preparation

There are overall 20,000 rows and 9 columns in the dataset, which means there are 20,000 respondents and 9 total variables. The 9 variables are written as “genhlth”,“exerany”,“hlthplan”,“smoke100”,“height”, “weight”, “wtdesire”,“age”,“gender”. And specifically, “genhlth” refers to the respondent’s general health, either as excellent, very good, good, fair, or poor. “exerany” represents the binary categorical data which answers whether the respondents exercised in the past month or not; “hlthplan” means the health plan, which indicates whether the respondents have health plan or not; “smoke100” indicates whether the respondent had smoked at least 100 cigarettes in their lifetime. “height” is measured in inches. “weight” is measured in pounds. “wtdesire” represents the respondent’s desired weight. “age” is measured in years. And for “gender” category, it is either as male or female. We performed data cleaning. First, we checked if missing values exist in the data. And it turns out that there are no missing values. But since our first research question wants to investigate weight loss, we would like to create a subset of the data where the respondent’s weight is larger than his desired weight. And we also need to add another new variable weightLoss,which takes the difference between weight and desired weight, to the data. The new data set is named as “cdc_new”. “cdc_new” has 12,764 rows, 10 columns.

EDA

EDA on Response Variable exerany:

Table 2.1 Number of Respondents Exercised in The Past Month

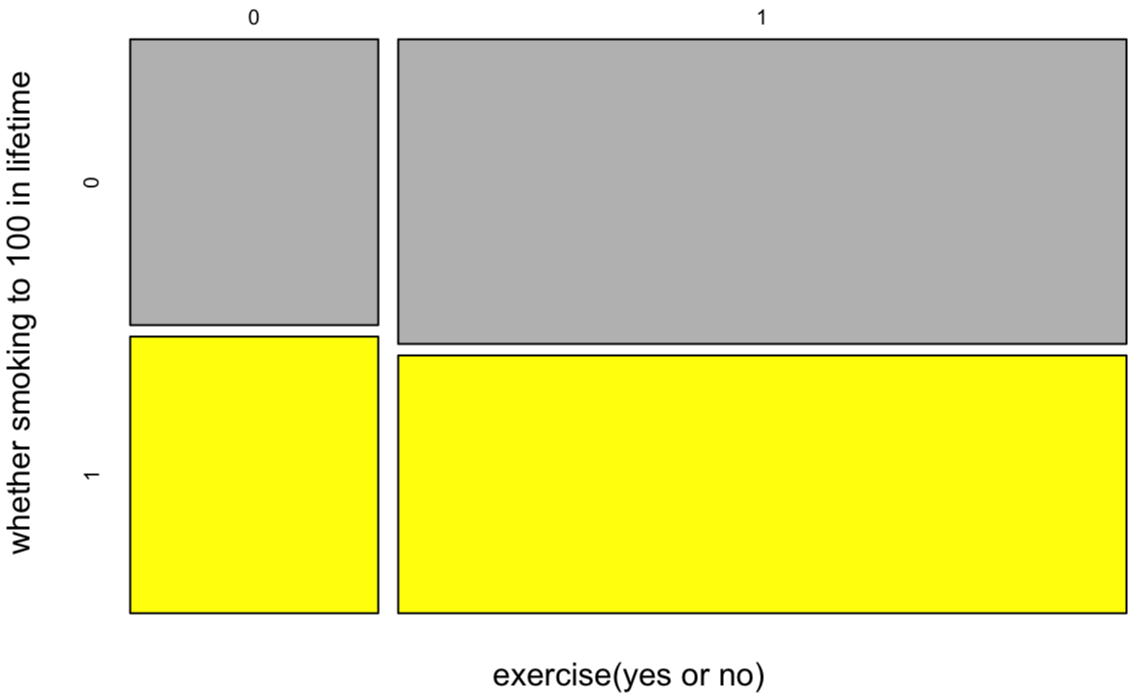
Var1	Freq
0	3243
1	9521

– Explanation on Table 2.1:

As shown from table 2.1, “exerany” is a categorical data which has two response : yes (1) or No(0). We have overall 9521 respondents who exercised in the past month, and 3243 respondents who did not exercise in the past month. In research question 1, we want to evaluate the relationship between the probability of the respondents exercising regularly and the weight they want to lose. Since “exerany” is a categorically binary data, we can interpret it in log odds form, and later construct the empirical logit plot to test the relationship between it and other variables.

EDA on smoke100:

Figure 2.2 The relationship between exercise regularly and smoke100



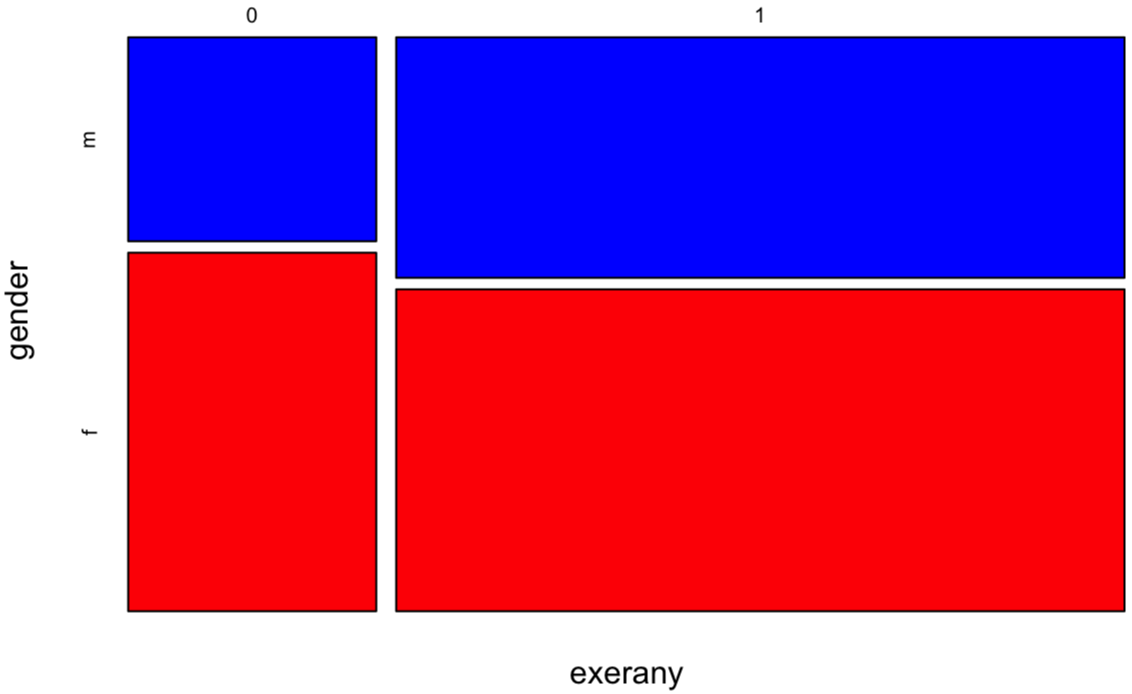
– Explanation on Figure 2.2:

smoke100 is one of the variables in the list. And we do EDA analysis here to prepare for the analysis in the second research question which needs consideration on other variables in the data. smoke100 is a categorical data so we will draw a mosaic plot to indicate its relationship with the probability of exercise.

As from Figure 2.2, we can somehow conclude that people who do not smoke up to 100 cigarettes in their lifetime has higher probability to exercise regularly. However, the relationship is not too obvious. So we will put a mark here and use hypothesis test to test the significance of this variable in improving the model accuracy.

EDA on gender:

Figure 2.3
The relationship between gender and exercise (yes or no)



– Explanation on Figure 2.3:

gender is another categorical variables in the list. Again, we will draw a mosaic plot to visualize the its relationship with gender. As shown from Figure2.3, among people who exercise regularly, there are more male than female. However, the difference is not large. So we will give a clearer look later when constructing the model.

EDA on height:

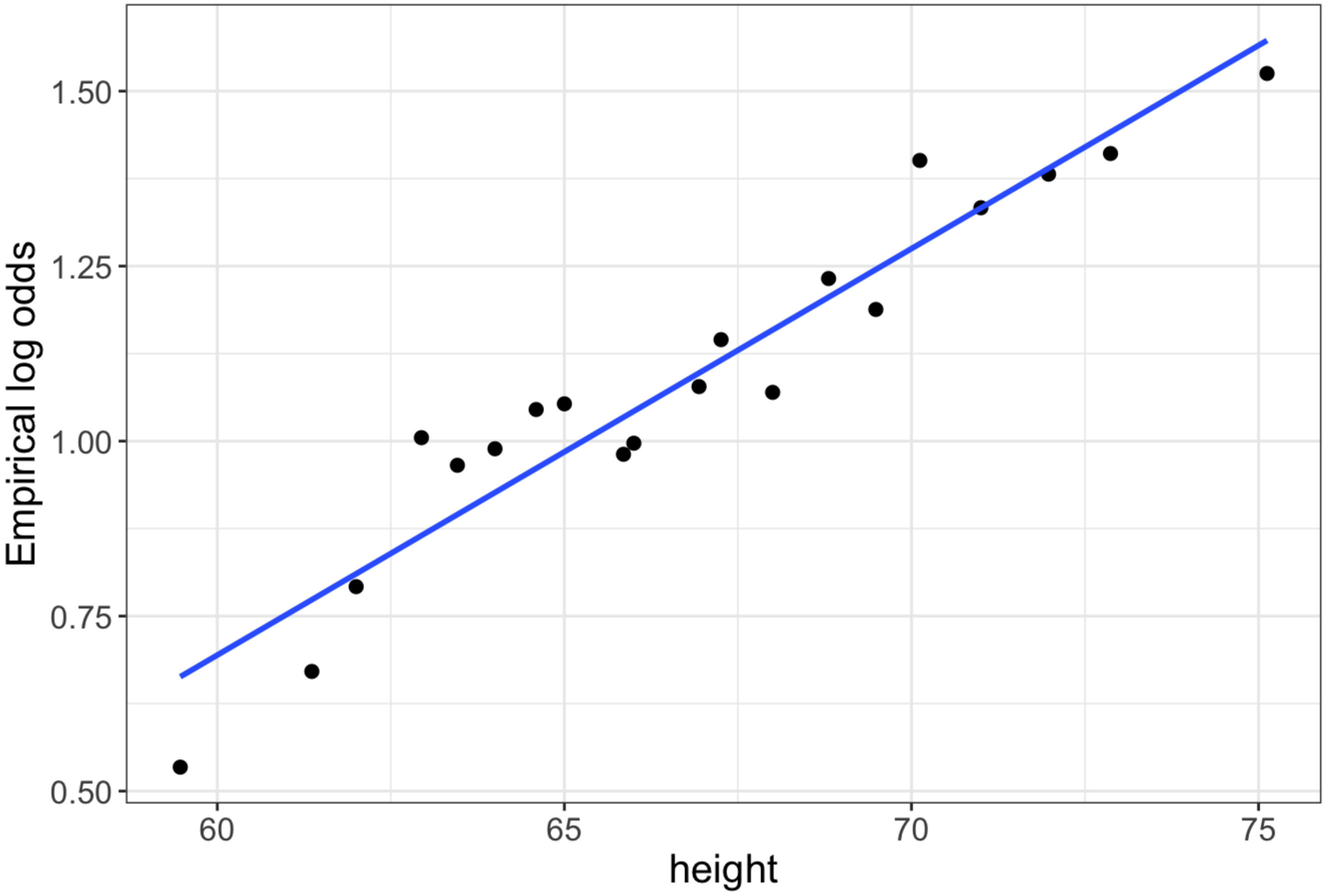


Figure 2.4 The relationship between log odds of exercising regularly and height

– Explanation on Figure 2.4:

The response variable is “exerany”, which had been analyzed previously as the categorical variable that has two responses. So we will choose an empirical logit plot to visualize its relationship with the variable height. As shown from Figure 2.4, height and log odds of “exerany” indicates a positive linear relationship. No further transformation is necessary.

EDA on age:

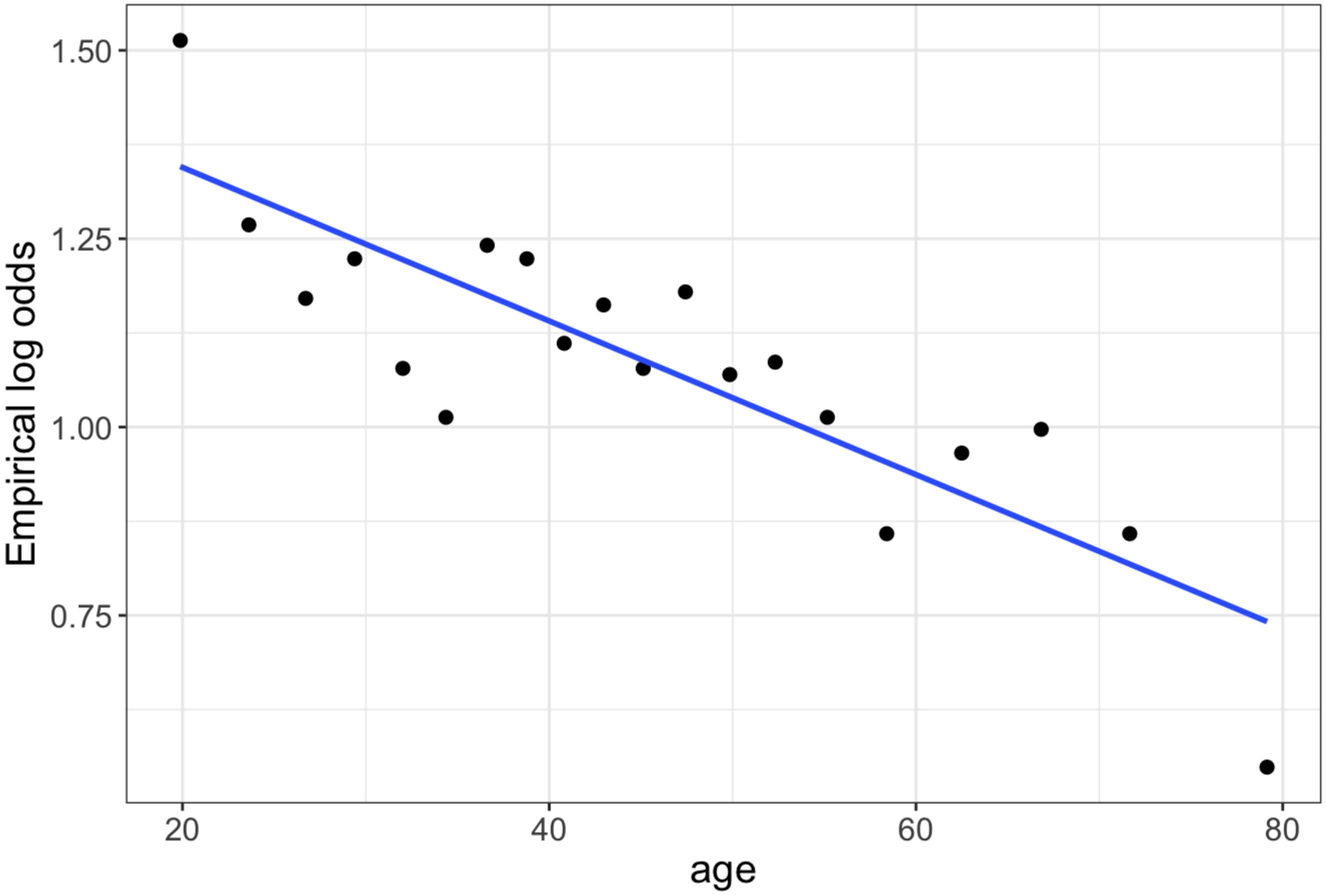


Figure 2.5 The relationship between the log odds of exercising regularly and age

– Explanation on Figure 2.5:

age is included in the already constructed model in research question 1. We performed EDA here just to figure out whether transformation on age is needed. As shown from Figure 2.5, a linear shape is a good fit for the model. So no further transformation should be applied.

EDA on weightLoss:

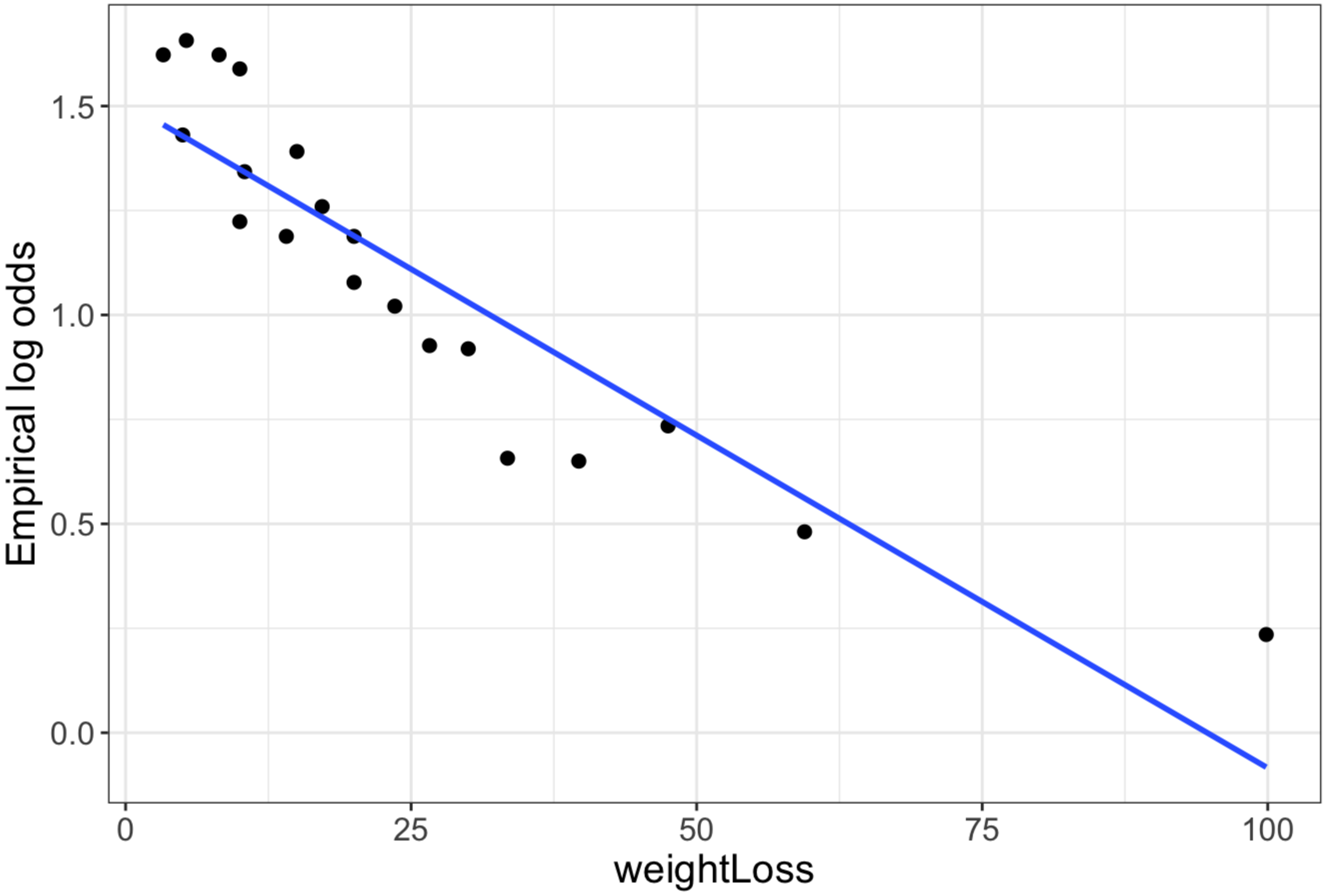


Figure 2.6 The relationship between log odds of exercising regularly and weightLoss

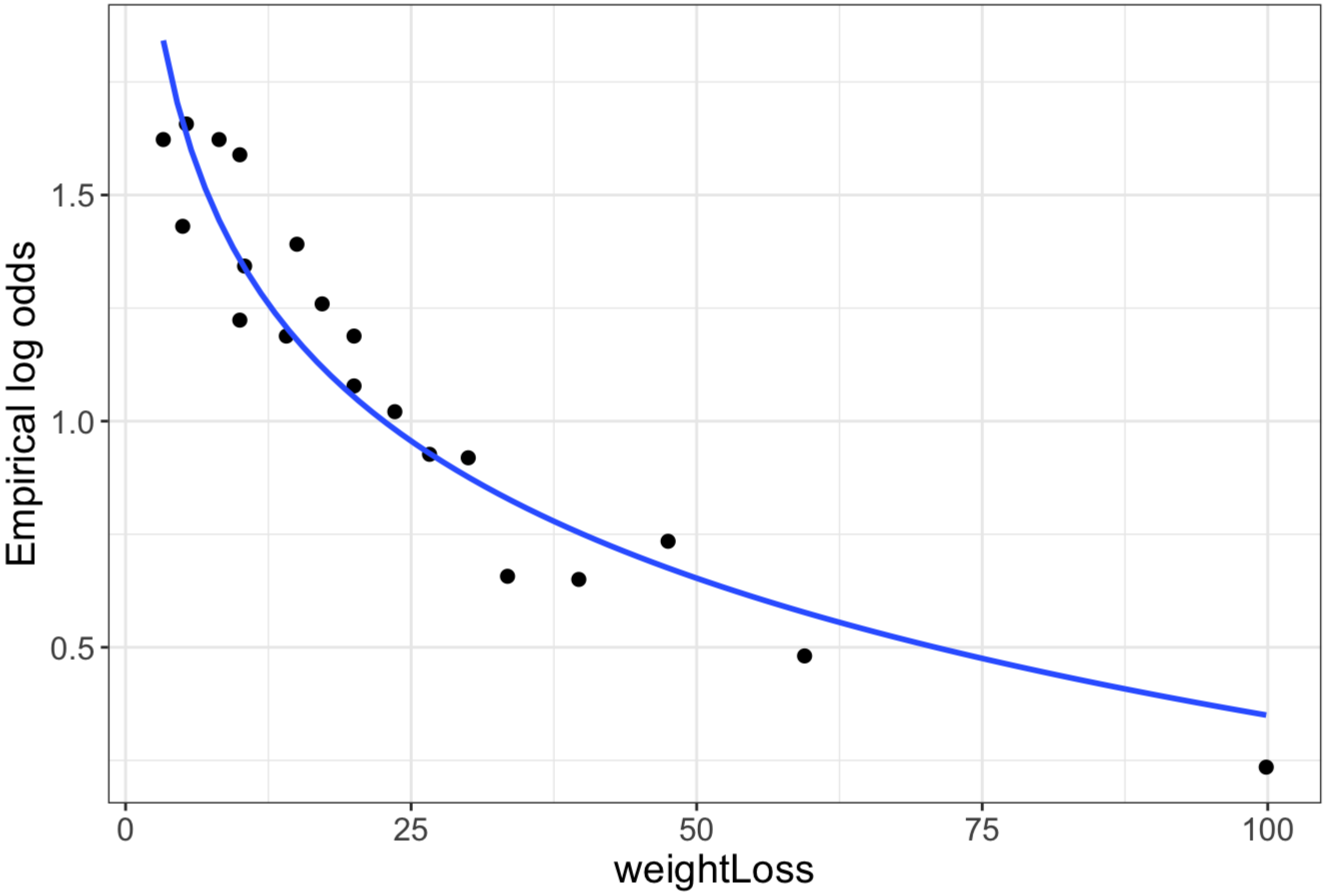


Figure 2.7 The relationship between the log odds of exercising regularly and log(weightLos

– Explanation on Figure 2.6 and Figure 2.7:

weightLoss is the explanatory variable we want to investigate in research question 1. Again, we will draw an empirical logit plot to visualize the relationship between weightLoss and log odds of “exerany”. From Figure 2.6, we found our that linear shape might not be a good fit for the data, so we applied further transformation from weightLoss to log(weightLoss). As shown from Figure2.7, log transformation seems like a wise choice. There is a negative relationship between log odds of “exerany” and log (weightLoss). 3

EDA on the potential Interaction Term of age and weightLoss:

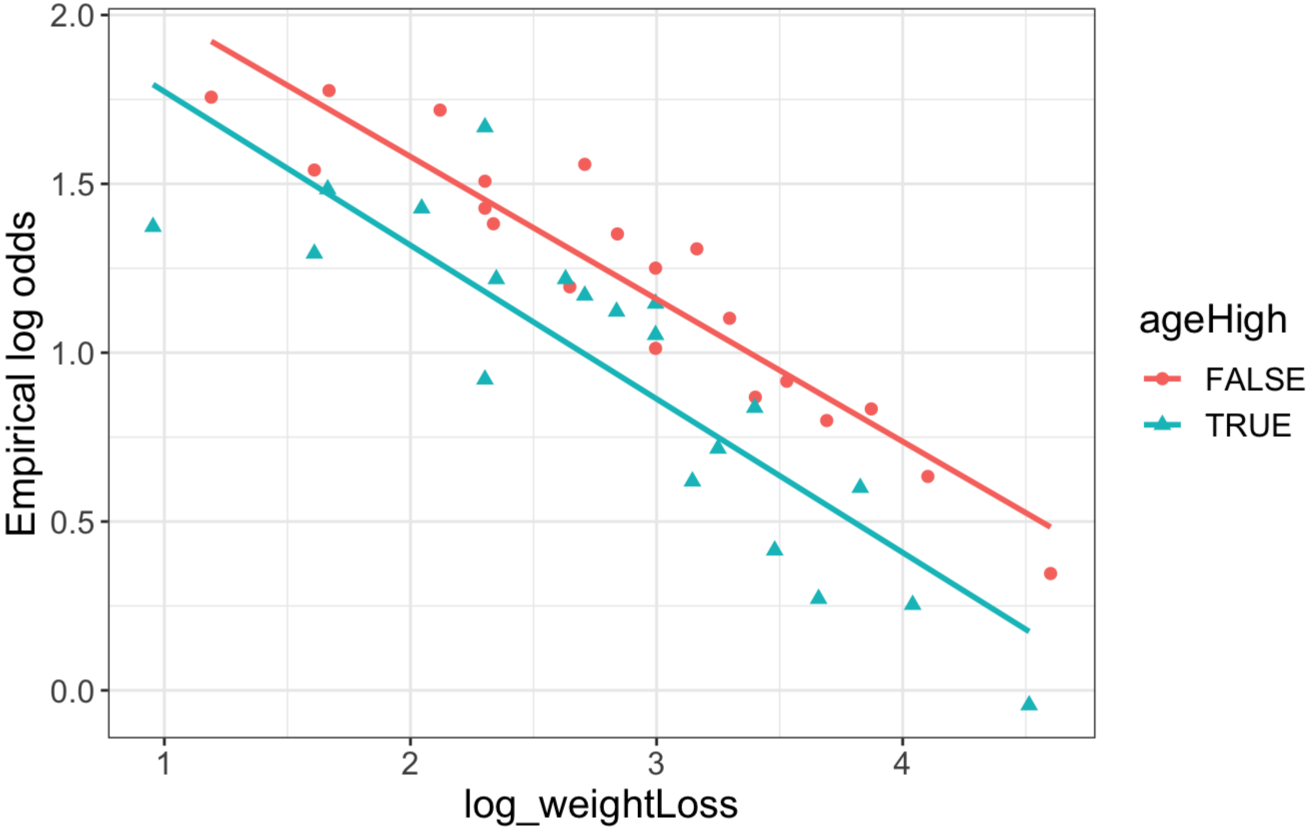


Figure 2.8
The relationship between log odds of exercising regularly and log(weightLoss), grouping by age

– Explanantion on Figure 2.8:

Since we need to add new variable weightLoss into model with existing variables age, general health, and health coverage. And it is reasonable to guess that there might be some relationship between respondent's age and the expected weightLoss. So we group the data by ages. We do it by setting the bound in age to 50, and therefore turn age into a categorical data.

As indicated by Figure2.8, two lines in the graph is almost parallel to each other. So no interaction term between weightLoss and age is needed when adding the variables.

EDA on the potential Interaction Term of genhlth and weightLoss:

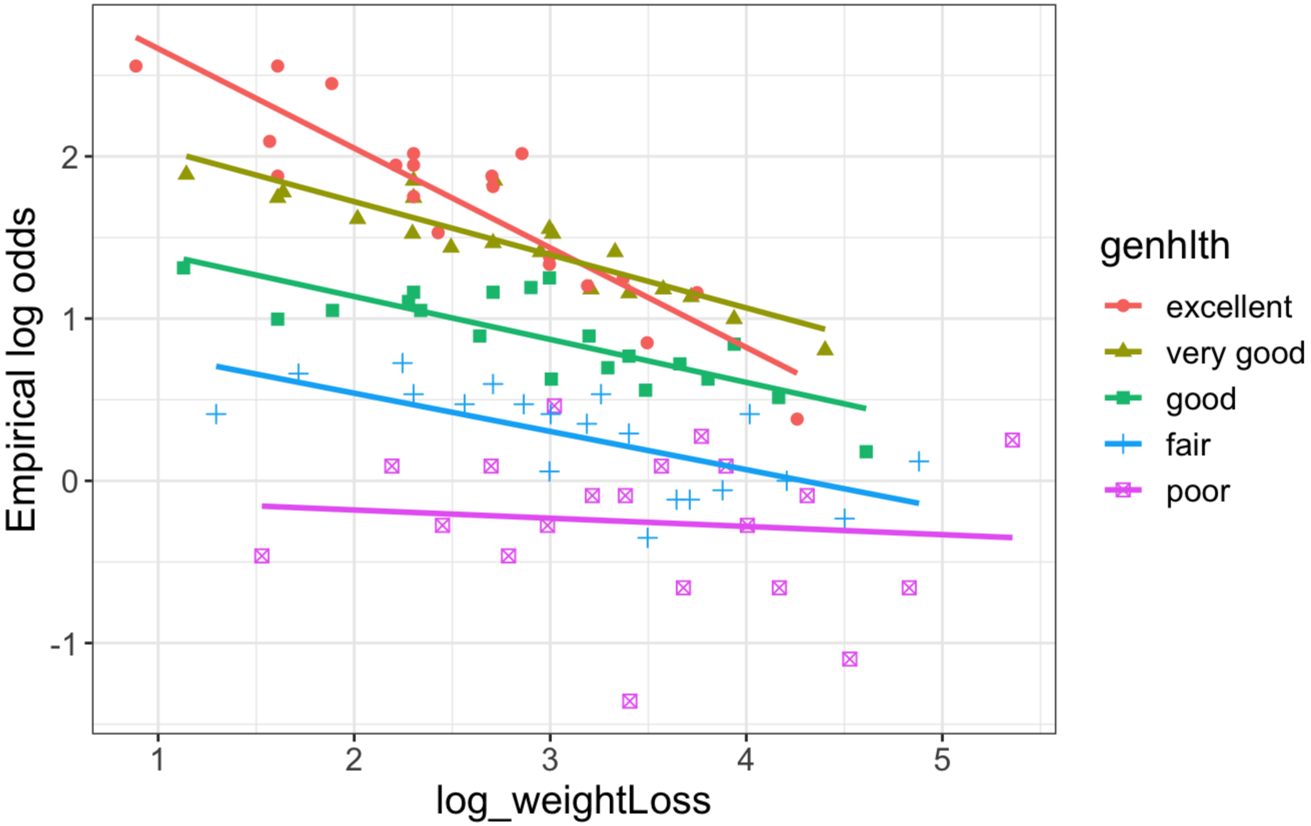


Figure 2.9
The relationship between log odds of exercising regularly and log(weightLoss), grouping by general health

– Explanantion on Figure 2.9:

As shown from Figure2.9, the lines which represent different health conditions in the graph intersects with each other, which indicates a correlation between weightLoss and general health. So later when constructing the model, we will add the interaction term of log(weightLoss) : genhlth into the model.

EDA on the potential Interaction Term of hlthplan and weightLoss:

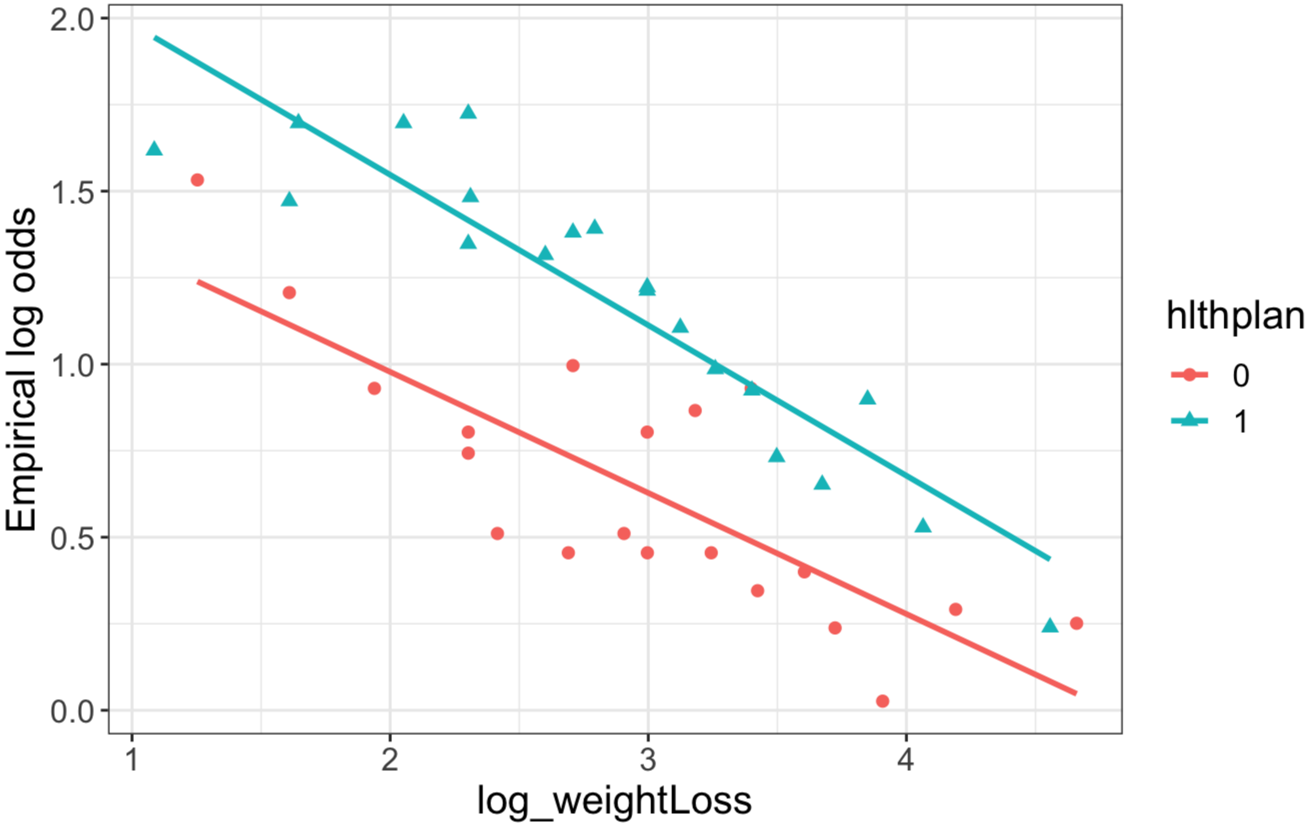


Figure 2.10
The relationship between log odds of exercising regularly
and log(weightLoss), grouping by health plan

– Explanation on Figure 2.10:

We also want to test the potential relationship between weightLoss and hlthplan. As shown from Figure 2.10, the lines is almost parallel to each other. Therefore, no interaction term is needed between these two variables.

EDA Summary:

exerany is a categorical data with two responses, and the relationship of it with other quantitative variables can be visualized using empirical logit plot. We also find that people who do not smoke up to 100 cigarettes in their lifetime has higher probability to exercise regularly.And among people who exercise regularly, there are more male than female. There is a positive relationship between height and log odds of exerany, and a negative relationship between log odds of exerany and log (weightLoss). Finally, we performed interaction term analysis and figured out that we should consider the interaction term of log(weightLoss) : genhlth when fitting the model.

Section 3: Modeling

The probability of the respondent exercising regugarly follows a Bernoulli distribution. We will then fit the parameter model in log odds form.

$$Y_i \sim Bernoulli(\pi_i)$$

Section 3.1: Exercise and weight loss goals

We will name the model with variables age, general health, and health coverage as model1, and the model with added new variable log(weightLoss) as model2. And since we figured out in EDA that there should be an interaction term between weightLoss and genhlth. We will construct model3 and perform hypothesis test to re-test the significance of the interaction term.

The three population models are given below:

$$model1 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 age + \beta_2 genhlth + \beta_2 hlthplan$$

$$model2 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 age + \beta_2 genhlth + \beta_2 hlthplan + \beta_3 log(weightLoss)$$

$$\begin{gathered} model3 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 age + \beta_2 genhlthverygood + \beta_3 genhlthgood + \\ \beta_4 genhlthfair + \beta_5 genhlthpoor + \beta_6 hlthplan + \beta_7 log(weightLoss) + \\ \beta_8 log(weightLoss) : genhlthverygood + \beta_9 log(weightLoss) : genhlthgood + \\ \beta_{10} log(weightLoss) : genhlthfair + \beta_{11} log(weightLoss) : genhlthpoor \end{gathered}$$

And then we fitted the model and get the coeefficient values for each parameters.

$$\begin{gathered} model1 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = 1.503862 - 0.006516 age - 0.176413 genhlthverygood \\ - 0.710101 genhlthgood - 1.273136 genhlthfair - 1.788470 genhlthpoor + 0.449996 hlthplan \end{gathered}$$

$$\begin{gathered} model2 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = 1.7662461 - 0.0080223 age - 0.1251584 genhlthverygood \\ - 0.6004535 genhlthgood - 1.09802466 genhlthfair - 1.5238098 genhlthpoor \\ + 0.4616550 hlthplan - 0.0109036 weightLoss \end{gathered}$$

$$\begin{gathered} model3 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = 3.298348 - 0.007768 age - 0.961609 genhlthverygood \\ - 1.597624 genhlthgood - 2.148762 genhlthfair - 3.021240 genhlthpoor + 0.443268 hlthplan \\ - 0.642136 log(weightLoss) + 0.309989 genhlthverygood : log(weightLoss) + \\ 0.362153 genhlthgood : log(weightLoss) + 0.375322 genhlthfair : log(weightLoss) + \\ 0.499941 genhlthpoor : log(weightLoss) \end{gathered}$$

1. Hypothesis Test on model1 and model2:

$$\begin{gathered} H0 : \beta_3 = 0 \\ Ha : \beta_3 \neq 0 \end{gathered}$$

– Evaluate the statistics:

The drop in deviance is 157.8678. We then use Chi square distribution to calculate out the probability of H0 to be true, which is 3.30767e-36. The probability is a really small number. We therefore have strong evidence that there is a relationship beteen log(weightLoss) and log odds of exercise regularly.

2. Hypothesis Test on model2 and model3:

$$\begin{gathered} H0 : \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0 \\ Ha : atleastoneof \beta_8/\beta_9/\beta_{10}/\beta_{11} \neq 0 \end{gathered}$$

– Evaluate the statistics

The drop in deviance is 46.27453. We then use Chi square distribution to calculate out the probability of H0 to be true, which is 2.159245e-09.

The probability is also a relatively small number. We therefore have strong evidence that there should be an interaction term beteen log(weightLoss) and general health.

Coefficient Interpretation

Therefore, we will choose model3 as our model to evaluate the relationship between the probability of exercise regularly and the expected weightLoss. Model3 is given below:

$$model3 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = 3.298348 - 0.007768 age - 0.961609 genhlthverygood \\ - 1.597624 genhlthgood - 2.148762 genhlthfair - 3.021240 genhlthpoor + 0.443268 hlthplan \\ - 0.642136 log(weightLoss) + 0.309989 genhlthverygood : log(weightLoss) + \\ 0.362153 genhlthgood : log(weightLoss) + 0.375322 genhlthfair : log(weightLoss) + \\ 0.499941 genhlthpoor : log(weightLoss)$$

age:

for one year increase in age, the log odds of exercising regularly decreases by 0.007768.

#####genhlthverygood: Compared to the respondents with excellent general health, respondents with gvery ood general health has 0.961609 smaller log odds of exercising regularly.

genhlthgood:

Compared to the respondents with excellent general health, respondents with good general health has 1.597624 smaller log odds of exercising regularly.

genhlthfair:

Compared to the respondents with excellent general health, respondents with good general health has 2.148762 smaller log odds of exercising regularly.

genhlthpoor:

Compared to the respondents with excellent general health, respondents with good general health has 3.021240 smaller log odds of exercising regularly.

hlthplan:

Compared to the respondents with no health coverage, respondents with health coverage has 0.443268 higher log odds of exercising regularly.

To make easier interpretation on log weightLoss, we will let weightLoss to be ‘e’ every unit.

log(weightLoss):

for every unit increase in weightLoss, the log odds of exercising regularly decreases by 0.642136.

log(weightLoss):genhlthverygood:

Compared with respondents who has excellent general health, for every one unit increase in weightLoss, the log odds of exercising regularly in respondents with very good health general health decreases by 0.642136 -0.309989 = 0.332147.

#### log(weightLoss):genhlthverygood:

Compared with respondents who has excellent general health, for every one unit increase in weightLoss, the log odds of exercising regularly in respondents with good health general health decreases by 0.642136 - 0.362153 = 0.279983.

#### log(weightLoss):genhlthveryfair:

Compared with respondents who has excellent general health, for every one unit increase in weightLoss, the log odds of exercising regularly in respondents with fair health general health decreases by 0.642136 - 0.375322= 0.266814.

#####log(weightLoss):genhlthverypoor: Compared with respondents who has excellent general health, for every one unit increase in weightLoss, the log odds of exercising regularly in respondents with poor health general health decreases by 0.642136 - 0.499941= 0.142195.

Modeling Conclusion

There is a generally negative relationship between log odds of exercising regularly and log(weightLoss), indicated by the negative coefficient -0.64236 in model3, And part of the negative value is offset by the interaction term of log(weigthLoss) and general health.

Section 3.2: Predicting exercise

There are three variables in the data that are left to be analyzed : smoke100, height, gender. As indicated from EDA, there is a positive linear relationship between height and log odds of exercising regularly. And we are not so certain on the significance of other two variables smoke100 and gender. So in this section, we will build three models, with first model adding new variables height to model3, second model adding smoke100, and third model adding gender.

Three models to testify the significance of variables height, smoke100, and gender.

$$model4 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 age + \beta_2 genhlthverygood + \beta_3 genhlthgood + \\ \beta_4 genhlthfair + \beta_5 genhlthpoor + \beta_6 hlthplan + \beta_7 log(weightLoss) + \\ \beta_8 log(weightLoss) : genhlthverygood + \beta_9 log(weightLoss) : genhlthgood + \\ \beta_{10} log(weightLoss) : genhlthfair + \beta_{11} log(weightLoss) : genhlthpoor + \beta_{12} height$$

$$\begin{gathered} model5 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 age + \beta_2 genhlthverygood + \beta_3 genhlthgood + \\ \beta_4 genhlthfair + \beta_5 genhlthpoor + \beta_6 hlthplan + \beta_7 log(weightLoss) + \\ \beta_8 log(weightLoss) : genhlthverygood + \beta_9 log(weightLoss) : genhlthgood + \\ \beta_{10} log(weightLoss) : genhlthfair + \beta_{11} log(weightLoss) : genhlthpoor + beta_{12} height + \beta_{13} smoke100 \end{gathered}$$

$$\begin{gathered} model6 : \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 age + \beta_2 genhlthverygood + \beta_3 genhlthgood + \\ \beta_4 genhlthfair + \beta_5 genhlthpoor + \beta_6 hlthplan + \beta_7 log(weightLoss) + \\ \beta_8 log(weightLoss) : genhlthverygood + \beta_9 log(weightLoss) : genhlthgood + \\ \beta_{10} log(weightLoss) : genhlthfair + \beta_{11} log(weightLoss) : genhlthpoor + \beta_{12} height + \beta_{13} gender \end{gathered}$$

Step1:

We will first give a hypothesis test between model3 and model4 to evaluate the significance of variable height. The hypothesis test is :

$$\begin{gathered} H0 : \beta_{12} = 0 \\ Ha : \beta_{12} \neq 0 \end{gathered}$$

The AIC of model3 is 13582.7, and AIC of model4 is 13492.02. The probability given by the hypothesis test is 6.157294e-22, whcih is a very small number. Therefore, we have stron gevidence that there is a relationship between log odds of exercising regularly and height. Height is then a statistically significant variable.

Step2:

Then, we can perform hypothesis test on model 4 and model 5 to test the significance of variable smoke100. model4 has AIC value of 13492.02, model5 has AIC value of 13493.82. Model4 is preferred because we will always choose model with the lower AIC.

And we can also perform hypothesis test on model4 and model5. The hypothesis test is:

$$\begin{gathered} H0 : \beta_{13} = 0 \\ Ha : \beta_{13} \neq 0 \end{gathered}$$

The probability of 0.655582, which is a large number. Therefore, we will keep the choice of model4.

Step3

: And finally, we will test the significance of variable gender by adding this variable to model6.

The AIC value of model6 is 13491.99, which is just slightly smaller than model4. And then we perform hypothesis test:

$$H0 : \beta_{13} = 0$$
$$Ha : \beta_{13} \ \neq 0$$

The probability of 0.1540326, which is not significant enough. therefore, we still keep our choice with model4.

Conclusion on the final model

The final model chosen is model4. From the model coeeficient, it is clear that people with better general health, less expected weight loss and higher height will have higher probability to exercise regularly.

$$finalmodel : \log\left(\frac{\pi_i}{1-\pi_i}\right) = -0.110014 - 0.006912 age - 0.938841 genhlthverygood$$
$$- 1.547124 genhlthgood - 2.090117 genhlthfair - 2.935558 genhlthpoor + 0.408704 hlthplan - 0.641348 log(weightLoss) +$$
$$0.302430 log(weightLoss) : genhlthverygood + 0.302430 log(weightLoss) : genhlthgood + 0.346535 log(weightLoss) : genhlthfair +$$
$$0.367600 log(weightLoss) : genhlthpoor + 0.050910 height$$

Predicting ability

Table 3.1: The Prediction Table

	0	1
0	373	297
1	2870	9224

Three metrics are used to evaluate the predicted ability of the model: overall accuracy, sensitivity and specificity. Sensitivity measures the true positive rate. It divides number of true positives by the sum of true positive and false negative . Specificity measures the true negative rate. It divides number of true negatives by the sum of true negatives and number of false positive.

The overall accuracy for the model is (373+9224) / (373+297+2870+9224) = 0.7519 The sensitivity for the model is (9224) / (9224+297) = 0.9688 The specificity for the model is (373) / (373+2870) = 0.1150

This model is great in predicting the overall accuracy and sensitity, but does not perform so well in specificity.

Section 4: Discussion

The probability of the respondents exercising regularly can be affected by several factors. It is negatively correlated with age, expected weight loss and positively related to general health conditions, health plan, and height. And to some extent, these variables are dependent on each other, which should also be considered as a statistically significant variable in the model.

It is clear from our model that there is a negative relationship between the probability of exercising regularly and expected weight loss, which specifically, expected weight loss is transformed into log forms in the model. And some part of its negativity is offset by the positive bring by adding the interaction term between expected weight loss and general health conditions.

The model we fit is good in the overall accuracy and sensitivity test, but not as well as in the specificity test. So we can consider putting in more other variables from BFRSS, for example, immunization, chronic health conditions, into the model to solve the limitation.