# Homework 06: Candy Competition

## Inference for Multiple Linear Regression

❗ Important

Due: Friday, October 20, 11:59pm

### Introduction

In today's homework you will analyze data about candy that was collected from an online experiment conducted at FiveThirtyEight.

#### Learning goals

By the end of the homework you will be able to

- Fit a linear model with multiple predictors and an interaction term
- Fit a linear model with categorical predictors
- Conduct inference on multiple linear models

### Getting started

- Go to RStudio and login with your College of Idaho Email and Password.
- Make a subfolder in your `hw` directory to store this homework.
- Log into Canvas, navigate to Homework 6 and upload the `hw-06.qmd` file into the folder your just made.

## Packages

The following packages are used in the lab.

```r
library(tidyverse)
library(broom)
library(ggformula)
library(fivethirtyeight)
library(knitr)
library(yardstick)

# add other packages as needed
```

## Data: Candy

The data from this lab comes from the the article FiveThirtyEight *The Ultimate Halloween Candy Power Ranking* by Walt Hickey. To collect data, Hickey and collaborators at FiveThirtyEight set up an experiment people could vote on a series of randomly generated candy matchups (e.g. Reeses vs. Skittles). Click here to check out some of the match ups.

The data set contains the characteristics and win percentage from 85 candies in the experiment. The variables are

| Variable | Description |
|---|---|
| `chocolate` | Does it contain chocolate? |
| `fruity` | Is it fruit flavored? |
| `caramel` | Is there caramel in the candy? |
| `peanutalmondy` | Does it contain peanuts, peanut butter or almonds? |
| `nougat` | Does it contain nougat? |
| `crispedricewafer` | Does it contain crisped rice, wafers, or a cookie component? |
| `hard` | Is it a hard candy? |
| `bar` | Is it a candy bar? |
| `pluribus` | Is it one of many candies in a bag or box? |
| `sugarpercent` | The percentile of sugar it falls under within the data set. Values 0 - 1. |
| `pricepercent` | The unit price percentile compared to the rest of the set. Values 0 - 1. |
| `winpercent` | The overall win percentage according to 269,000 matchups. Values 0 - 100. |

Use the code below to get a glimpse of the `candy_rankings` data frame in the **fivethirtyeight** R package.

```r
glimpse(candy_rankings)
```

```
Rows: 85
Columns: 13
$ competitorname   <chr> "100 Grand", "3 Musketeers", "One dime", "One quarter~
$ chocolate        <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, FALSE, F~
$ fruity           <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE~
$ caramel          <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE,~
$ peanutyalmondy   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, ~
$ nougat           <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE,~
$ crispedricewafer <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE~
$ hard             <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
$ bar              <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, FALSE, F~
$ pluribus         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE~
$ sugarpercent     <dbl> 0.732, 0.604, 0.011, 0.011, 0.906, 0.465, 0.604, 0.31~
$ pricepercent     <dbl> 0.860, 0.511, 0.116, 0.511, 0.511, 0.767, 0.767, 0.51~
$ winpercent       <dbl> 66.97173, 67.60294, 32.26109, 46.11650, 52.34146, 50.~
```

## Exercises

The goal of this analysis is to use multiple linear regression to understand the factors that make a good candy.

### Exercise 1

Notice that the values of `pricepercent` and `sugarpercent` are proportions. Change the scale so that they are percentages.

### Exercise 2

- Our response variable in this homework will be `winpercentage`. Choose two additional variables, one quantitative and one categorical. Generate a SINGLE plot that visualizes all three variables. Hint: remember that you can tie any aesthetic in your plot (e.g. color) to a variable be writing `aesthatic = ~variable_name`.
- Write two observations from your plot.

### Exercise 3

Fit a linear model including both variables you chose above and include an interaction term between the two.

### Exercise 4

Interpret the following in the context of the data:

- Intercept
- Coefficient of your quantitative variable
- Coefficient(s) of your categorical variables
- Coefficient(s) of your interaction term.

### Exercise 5

Choose one of the coefficients from your model and write out all the steps in the hypothesis test that corresponds to it's p-value as in this slide.

### Exercise 6

Interpret the other p-values in your model in the context of the data. You do not need to write out the full hypothesis testing framework as you did in Exercise 5.

### Exercise 7

Generate 95% confidence intervals for all of the slopes in your model and interpret them in the context of the data. Are these intervals independent of one another?

### Exercise 8

Fit a linear model predicting `winpercent` using `chocolate` and `pricepercent` as predictors. Include a interaction term between `pricepercent` and `chocolate`. Describe the effect of `pricepercent` for chocolate candy in the context of the data.

## Exercise 9

- Consider the model from Exercise 3 “Model 1” and the model fit in Exercise 8 “Model 2”. If you happened to have chosen `pricepercent` and `chocolate` in Exercise 3, please choose two new variables and to complete this problem with.
- Which model would you choose based on $R^2$? Briefly explain your choice.
- Which model would you choose based on $RMSE$? Briefly explain your choice.

## Exercise 10

- Use the model you selected in Exercise 9 to describe what generally makes a good candy, i.e., one with a high win percentage.

## Grading

Total points available: 50 points.

| Component | Points |
|---|---|
| Ex 1 | 2 |
| Ex 2 | 3 |
| Ex 3 | 5 |
| Ex 4 | 8 |
| Ex 5 | 6 |
| Ex 6 | 5 |
| Ex 7 | 5 |
| Ex 8 | 4 |
| Ex 9 | 4 |
| Ex 10 | 4 |
| Workflow & formatting | 4[1] |

[1] The “Workflow & formatting” grade is to assess the reproducible workflow, clarity, and professionalism.