Lesson 18 Cereal

LTC J. K. Starling

Last compiled on 02 March, 2022

Background Added sugar may very well be the single worst ingredient in the modern diet. It contributes to several chronic diseases, and most people are eating way too much of it. Notably, most of this sugar comes from processed foods — and breakfast cereals are among the most popular processed foods that are high in added sugars. In fact, most cereals list sugar as the second or third ingredient. Starting the day with a high-sugar breakfast cereal will spike your blood sugar and insulin levels. A few hours later, your blood sugar may crash, and your body will crave another high-carb meal or snack — potentially creating a vicious cycle of overeating. Excess consumption of sugar may also increase your risk of type 2 diabetes, heart disease, and cancer.

Today we will take a look at the association between sugar, calories, and how people rate popular breakfast cereals.

```
# Load packages
library(tidyverse)
library(ggResidpanel)
library(car)

### Data management
# cereal <- read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/cereal.csv", header = TRU

# setwd("C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons/Block III/LSN18/")
# cereal <- read.csv("cereal.csv", header = TRUE, stringsAsFactors = TRUE)

# select calories, sugar, and rating
# cereal <- cereal %>% select(...)
```

Single-variable models

(1) Ok. Let us look at the one-variable models. Which explanatory variable, sugar or calories, has the largest impact on rating?

- What are we using to decide?

- Can we use the partial F test to decide?

- In what situation can we use the partial F-test?

```
# sugar.lm = lm(...
#
# calories.lm<-lm(...
```

- What do we conclude based on the R^2?

- What does it mean that the sum of the R^2 for the two individual models is greater than one?

Two-variable model

(2) What is the mathematical model that helps us answer the research question?

(3) What is a key advantage to analyzing the two variables simultaneously?

```
# Figure 5.1.5: Two-variable regression model fits a plane of predicted values (cereal rating)

# both.lm <- lm(...
```

(4) What are our conclusion based on the p-values? How do we interpret coefficients?

Key idea: If you have a balanced, factorial design, even with quantitative explanatory variables, the slope coefficients in the two-variable model will be the same as in the one-variable models because you have designed the study so that there is no association between the explanatory variables. However, this is not the case in this example since we have an observational study.

Two-variable model with standardized factors

(5) Can we look at the $\hat{\beta}_1$ and $\hat{\beta}_2$ and deduce which factor has more of an effect on predicted rating? Why or why not?

- What do we need to do to the factors so that we can look directly at the magnitude of the slope coefficients to determine which variable has more of an effect on percentage peroxide.

- How do we standardize?

- Will this mess up the strength of the associations with the response?

- What is added advantage of standardizing quantitative variables?

- Standardize the data. Provide a scatter plot of the regular and standardized data.

```
# cereal_std = cereal %>%
  # mutate_at(list(~scale(.)), .var = vars(....))

# plot regular and standardized data
# cereal %>% ggplot(...
```

- Fit the model with the standardized variables.

```
# std.lm<-lm(...

# must calculate to fully interpret coefficients
# cereal %>% select(grams.of.sugars, calories.per.serving) %>%
```

```
# summarise_all(list(~mean(.), ~sd(.)))
```

(6) What are our conclusion based on the p-values? How do we interpret coefficients?

Key idea: Remember, all relationships or associations are not correlations. The authors note that if we also standardized the response variable, then the standardized slope coefficients would be equal to the correlation coefficients for the adjusted variables.

(7) To see if this is a good model, we of course have to check the validity conditions. What are they? Do we have any concerns?

```
# resid_panel(..., plots = c("hist", "resid"))
```

Interaction model

(8) What does an interaction between two quantitative variables mean?

```
# inter.lm<-lm(...
```

Key idea: An additional advantage to standardizing variables is the interaction product variable will not be linearly related to either variable involved in the interaction. In other words, the coefficient of determination is additive.

How can we test whether a model with an interaction is preferable to the two-variable model?

```
# anova(...
```

(9) What are our conclusion based on the p-values?

(10) Let's check the validity conditions for the interaction model.

```
# resid_panel(..., plots = c("hist", "resid"))
```

(11) Let's plot the data in a 3D scatter-plot: