Lesson 18 Cereal (5.1)

LTC J. K. Starling

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)
library(cowplot)

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

# 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(calories.per.serving, grams.of.sugars,Rating.of.cereal)
```

Single-variable models

1) Let us look at the single-variable models.

```
sugar.lm <- cereal %>% lm(Rating.of.cereal ~ grams.of.sugars, data = .)
summary(sugar.lm)
```

```
##
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.853  -5.677  -1.439   5.160  34.421
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      59.2844     1.9485   30.43  < 2e-16 ***
## grams.of.sugars  -2.4008     0.2373  -10.12 1.15e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.196 on 75 degrees of freedom
## Multiple R-squared:  0.5771, Adjusted R-squared:  0.5715
## F-statistic: 102.3 on 1 and 75 DF,  p-value: 1.153e-15
```

```
calories.lm <- cereal %>% lm(Rating.of.cereal ~ calories.per.serving, data = .)
summary(calories.lm)
```

```
##
## Call:
## lm(formula = Rating.of.cereal ~ calories.per.serving, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.7201  -7.9317  -0.6678   5.9902  23.4161
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          95.78802    6.55057  14.623  < 2e-16 ***
## calories.per.serving -0.49701    0.06031  -8.241 4.14e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.24 on 75 degrees of freedom
## Multiple R-squared:  0.4752, Adjusted R-squared:  0.4682
## F-statistic: 67.92 on 1 and 75 DF,  p-value: 4.14e-12
```

2) What are the single-variable models? Write them down here.

3) Which explanatory variable, sugar or calories, has the largest impact on rating? What are we using to decide?

4) Can we use the partial F test to decide?

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

7) What do we conclude based on the R^2 values?

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

Two-variable model

9) What is the mathematical model that helps us answer the research question? Write it down.

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

11) Consider the multiple regression model summary below. Is this a better model? What are our conclusion based on the p-values? How do we interpret coefficients?

```
both.lm <- cereal %>% lm(Rating.of.cereal ~ grams.of.sugars + calories.per.serving, data = .)
summary(both.lm)
```

```
##
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars + calories.per.serving,
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.643  -6.339  -1.221   4.823  23.413
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          84.11417    5.44513  15.448  < 2e-16 ***
## grams.of.sugars      -1.71939    0.25225  -6.816 2.16e-09 ***
## calories.per.serving -0.27644    0.05755  -4.804 7.93e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.083 on 74 degrees of freedom
## Multiple R-squared:  0.6776, Adjusted R-squared:  0.6689
## F-statistic: 77.78 on 2 and 74 DF,  p-value: < 2.2e-16
```

12) **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.** Why is this not the case with this example data set?

Two-variable model with standardized factors

13) 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?

```
summary(cereal$calories.per.serving)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    50.0   100.0   110.0   106.9   110.0   160.0
```

```
summary(cereal$grams.of.sugars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.000   3.000   7.000   6.922  11.000  15.000
```

14) 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 cereal rating?

15) How do we standardize?

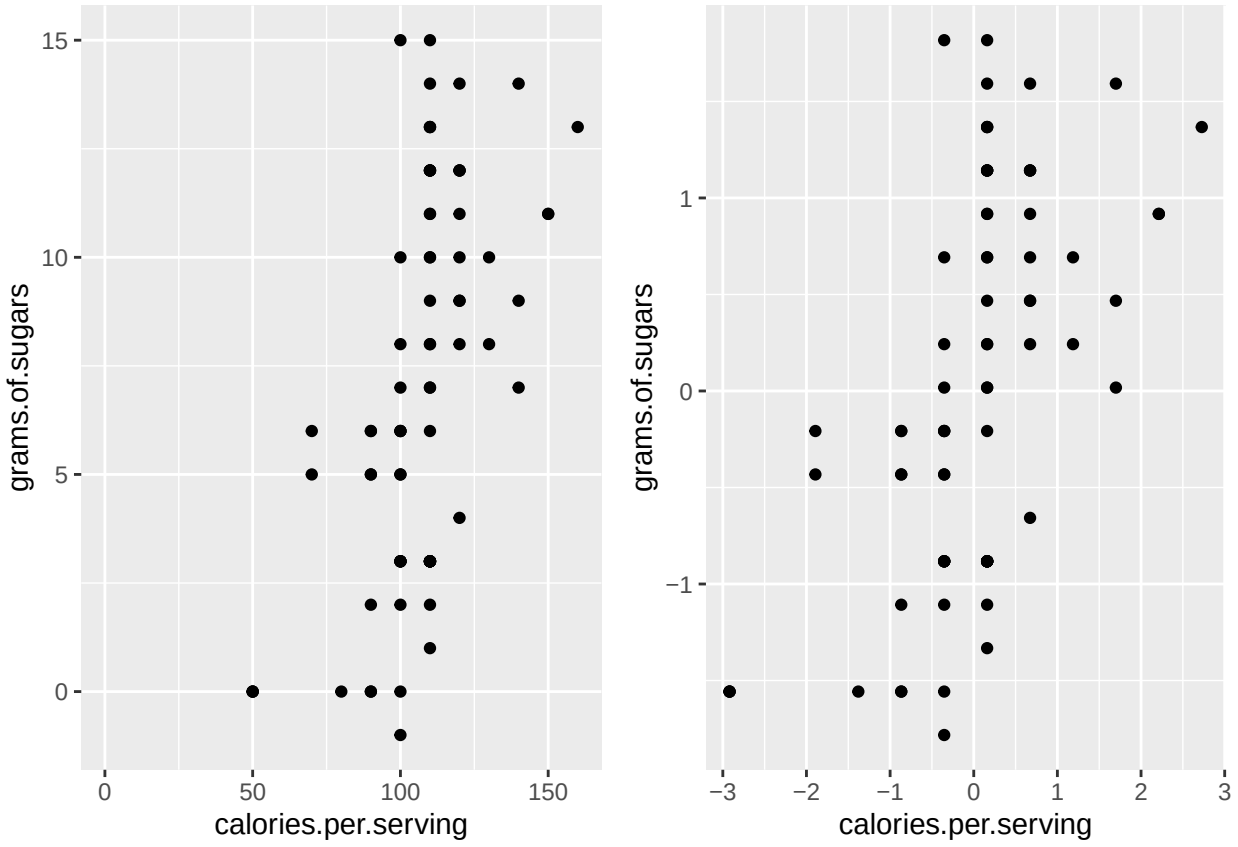
16) Standardize the data. Provide a scatter plot of the regular and standardized data. Explain what you see.

```
cereal_std = cereal %>%
  mutate_at(list(~scale(.)), .var = vars(grams.of.sugars, calories.per.serving))

# plot regular and standardized data
p.reg <- cereal %>% ggplot(aes(x=calories.per.serving, y=grams.of.sugars)) +
  geom_point() +
  expand_limits(x = 0, y = 0)
```

```
p.std <- cereal_std %>% ggplot(aes(x=calories.per.serving, y=grams.of.sugars)) +
  geom_point()

plot_grid(p.reg, p.std, ncol=2)
```



17) Will this mess up the strength of the associations with the response?

18) What is added advantage of standardizing quantitative variables?

19) Fit the model with the standardized variables.

```
std.lm <- cereal_std %>% lm(Rating.of.cereal ~ grams.of.sugars + calories.per.serving, data = .)
summary(std.lm)
```

```
##
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars + calories.per.serving,
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.643  -6.339  -1.221   4.823  23.413
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            42.6657     0.9211  46.320  < 2e-16 ***
## grams.of.sugars        -7.6425     1.1212  -6.816 2.16e-09 ***
## calories.per.serving   -5.3862     1.1212  -4.804 7.93e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.083 on 74 degrees of freedom
## Multiple R-squared:  0.6776, Adjusted R-squared:  0.6689
## F-statistic: 77.78 on 2 and 74 DF,  p-value: < 2.2e-16
```

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

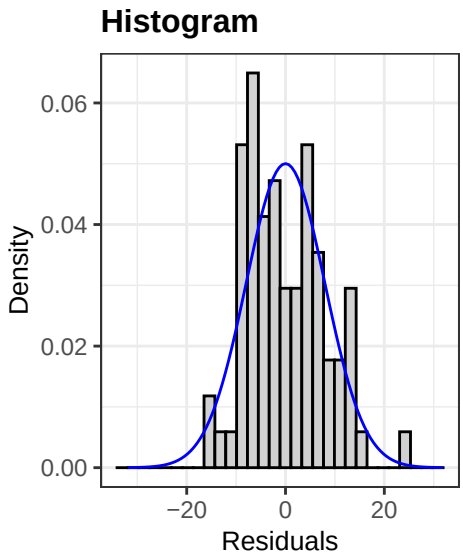
```
##   grams.of.sugars_mean calories.per.serving_mean grams.of.sugars_sd
## 1             6.922078                  106.8831           4.444885
##   calories.per.serving_sd
## 1                19.48412
```

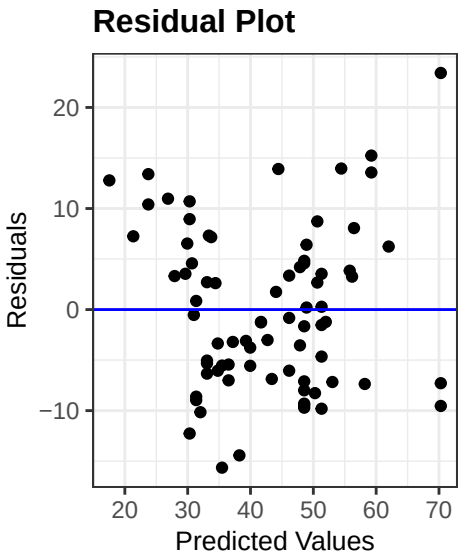
What are our conclusion based on the p-values? How do we interpret coefficients?

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.

20) 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(std.lm, plots = c("hist", "resid"))
```





Interaction model

21) Create the interaction model. What does an interaction between two quantitative variables mean?

```
inter.lm <- cereal_std %>% lm(Rating.of.cereal ~ grams.of.sugars * calories.per.serving, data = .)
summary(inter.lm)
```

```
##
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars * calories.per.serving,
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1761  -5.3793   0.1491   4.8486  15.7521
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         41.1729     0.9893  41.618  < 2e-16 ***
## grams.of.sugars                     -7.6436     1.0586  -7.221 4.05e-10 ***
## calories.per.serving                -4.6679     1.0826  -4.312 5.00e-05 ***
## grams.of.sugars:calories.per.serving  2.6896     0.8498   3.165  0.00226 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.631 on 73 degrees of freedom
## Multiple R-squared:  0.7165, Adjusted R-squared:  0.7049
## F-statistic: 61.51 on 3 and 73 DF,  p-value: < 2.2e-16
```

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.

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

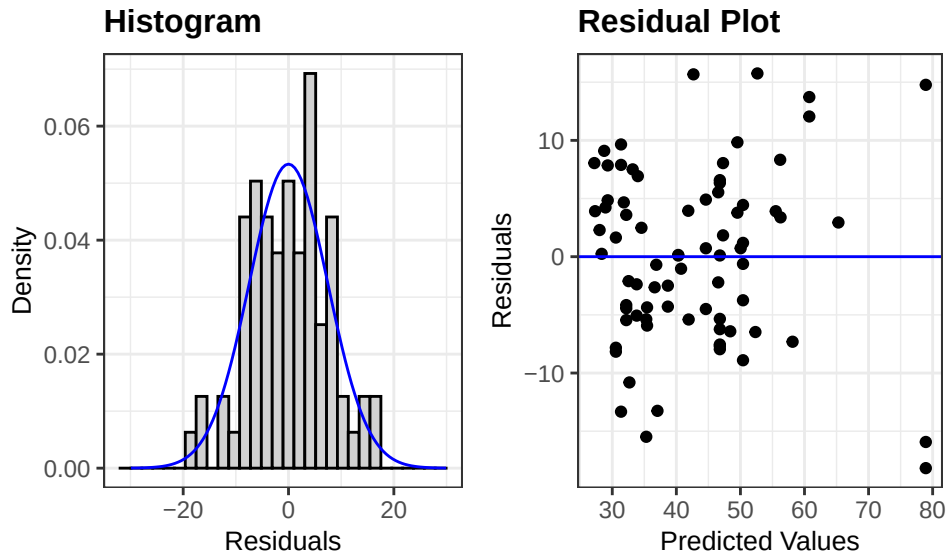
23) What are our conclusion based on the p-values?

```
anova(std.lm,inter.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Rating.of.cereal ~ grams.of.sugars + calories.per.serving
## Model 2: Rating.of.cereal ~ grams.of.sugars * calories.per.serving
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     74 4834.5
## 2     73 4251.1  1    583.39 10.018 0.002261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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



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