Lesson 18 Solution

MAJ Daniel Baller

Last compiled on 28 February, 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)
```

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?

Coefficient of determination.

- Can we use the partial F test to decide?

No. These models are not nested.

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

To decide whether the two-variable model is preferred over either of our single variable models.

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

```
## 
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars, data = cereal)
## 
## 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<-lm(Rating.of.cereal~calories.per.serving,data=cereal)
summary(calories.lm)
```

```
## 
## Call:
## lm(formula = Rating.of.cereal ~ calories.per.serving, data = cereal)
## 
## 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
```

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

The association between sugar and rating is stronger than the association between calories and rating.

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

There is some covariation between our two explanatory variables. i.e. some of the variation explained by sugar is jointly explained by calories.

Two-variable model

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

Multiple regression model

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

Reduces the unexplained variation.

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

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

```
##
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars + calories.per.serving,
##     data = cereal)
##
## 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
```

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

\textcolor{blue}{*The model that includes both variables is significantly better than using the single mean model. We are now accounting for 67.76% of the variation in ratings.*}

Sugar is significantly associated with rating. Predicted rating will decrease by 1.71 points with one gram increase in sugar per serving holding calories constant.

Calories is significantly associated with rating. Predicted rating will decrease by 0.21 points with one calorie increase per serving holding sugar constant.

The intercept of 84.11 is the predicted average rating when sugar and calories are both zero (nonsense!).

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?

The factors are on different scales so we cannot compare them directly.

- 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.

Standardize.

- How do we standardize?

To standardize a variable we subtract the mean and divide by the standard deviation. Each standardized variable will have a mean of zero and a standard deviation of one.

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

Standardizing the two variables has done nothing to change the strength of the associations with the response variable.

- What is added advantage of standardizing quantitative variables?

We can get a more meaningful intercept that is guaranteed to be in the middle of your data, rather than involving extrapolation or nonsensical values).

Ok. Let us fit the model with the standardized variables.

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

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

```
##
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars + calories.per.serving,
##     data = cereal_std)
##
## 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
```

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

The model that includes both variables is significantly better than using the single mean model. Note that this has not changed compared to the model with the unstandardized variables, which is what we expected.

The intercept is the predicted rating when sugar is 6.92 grams (its mean value) and when calories is 106.88 per serving (its mean value) OR when when both the standardized variables are 0

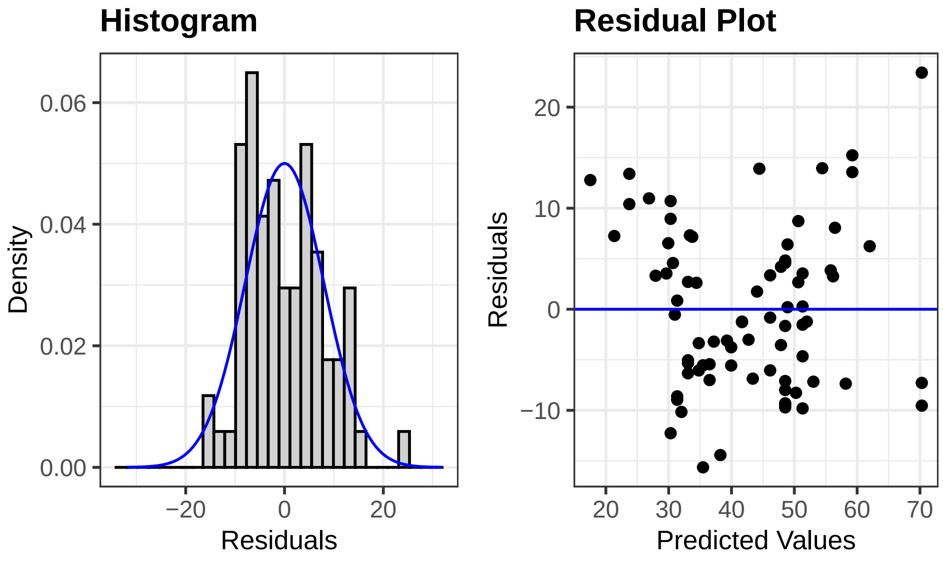
Sugar is significantly associated with rating (unchanged). Predicted rating will decrease by 7.62 points for each one standard deviation increase in sugar (about 4.44 grams) when calories per serving constant (calories per serving constant = calories per serving mean = 106.88 calories per serving). OR Predicted rating will decrease by 7.62 points for each unit increase in standardized sugar when holding standardized calories = 0.

Calories significantly associated with rating (unchanged). Predicted rating will decrease by 5.38 points with each standard deviation increase in calories per serving (about 19.48) when holding sugar constant (sugar constant = sugar mean = 6.92 grams of sugar). OR Predicted rating will decrease by 5.38 points for each unit increase in standardized calories when holding standardized sugar = 0.

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?

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



The residuals vs. predicted values plot does show some patterns to that suggest an interaction model may be appropriate.

Interaction model

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

The slope of one explanatory variable changes by a fixed amount as the other explanatory variable value changes.

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

```
##
## Call:
## lm(formula = Rating.of.cereal ~ grams.of.sugars * calories.per.serving,
##     data = cereal_std)
##
## 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.

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

Using the ANOVA (since these models are nested. We'll assume that the validity conditions are satisfied.).

```r
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
```

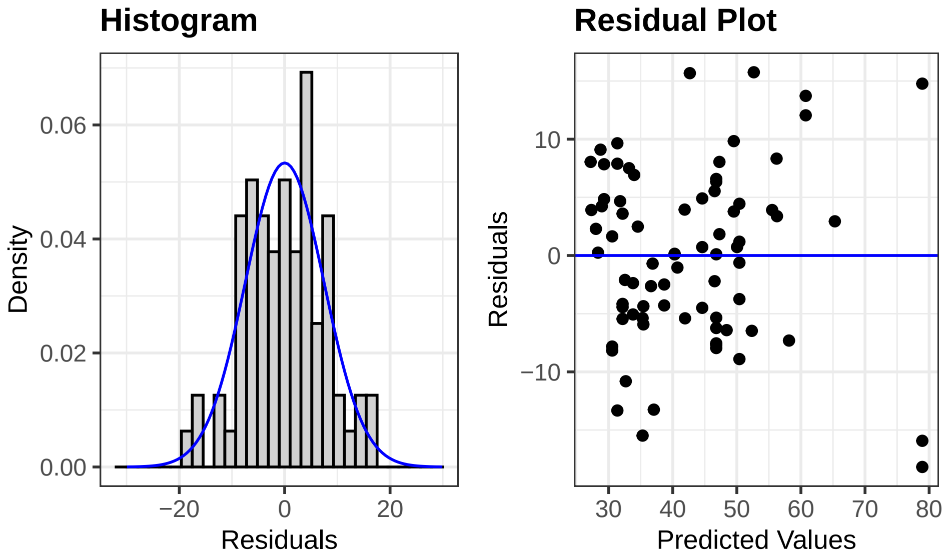
(9) What are our conclusion based on the p-values?

From the F-statistic and p-value we see that the interaction model is significantly better than the model without the interaction term.

From the F-statistic and p-value we see that the interaction model is significantly better than the model without the interaction term.

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

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


Histogram
Density
Residuals
Residual Plot
Residuals
Predicted Values