Lesson 16 Pizza.

by LTC J. K. Starling

Last compiled on 24 February, 2022

Background: A consumer organization rated various frozen pizzas on taste. The results are in the file Pizza.csv where the Rating is out of 100 (higher is better), the Calories are per serving, the Fat is in grams per serving, and the Pepperoni variable indicates whether there was pepperoni on the pizza (0-No; 1-Yes).

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

# pizza.dat <- read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/Pizza_w_pepp.csv", str
setwd("C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons/Block III/LSN16/")
pizza.dat <-read_csv("Pizza_w_pepp.csv")
# glimpse(pizza.dat)

# Change categorical variables to 'factors'
pizza.dat <- pizza.dat %>% mutate(Pepperoni = as.factor(Pepperoni))
```

Step 1: Ask a research question.

What factors are associated with pizza taste rating?

Step 2: Design a study and collect data.

(1) Is this study an observational study or a randomized experiment? Explain.

Step 3: Explore the data

(2) Create individual plots to explore the association between taste rating and each explanatory variables.

Step 4: Draw inferences beyond the data Multiple-Means Models

Up until now we have been using statistical models of the form:

$$\hat{y}_{ij} = \mu_j \quad \text{or} \quad \hat{y}_{ij} = \mu + \alpha_j$$

In the multiple means examples we have looked at, our explanatory variable, or independent variable have been categorical. This was nice in that we could think of each of our observations as belonging to a group - one level of the categorical variable. We considered whether there was an association between the response variable and the explanatory variable.

In this section, you will continue to focus on using one explanatory variable to explain variation in a response variable; however, the explanatory variable will be quantitative and we will consider whether the means of the response variable distributions at different values of the explanatory variable tend to follow a linear pattern (i.e., a constant rate of change).

But first… let's create a factor in the pizza.dat dataframe to account for those servings that are considered low-calorie. Label those servings with with less than or equal to 335 calories as 'low' and those greater than 335 calories as 'high'.

(3) Create a multiple-means model with `Lowcal` as the explanatory variable and `Rating` as the response variable. Show the summary and anova table.

```
# lowcal.lm <-
```

(4) Interpret your results.

Simple Linear Regression Models

Consider the following linear regression model (indicator coding).

$$y_i = \beta_0 + \beta_1 \mathrm{x}_1 + \epsilon_i \quad \epsilon_i \sim \mathrm{Normal}(0, \sigma^2)$$

- Interpret β_1 in the model.

- Interpret β_0 in the model.

Fit a simple linear regression model for predicting Rating taking into account the number of calories per serving (use indicator coding).

(5) Interpret the coefficient/estimate for Calories. Is there evidence of a significant association between Calories and Rating?

```
# calories.lm <-
```

(6) How does this model compare to the multiple-means model we made above?

(7) Are the validity conditions for the theory-based test satisfied?

- L

- I
- N
- E

```r
# resid_panel(
```

Fit a simple linear regression model for predicting Rating taking into account the amount of fat per serving (use indicator coding).

(8) Interpret the coefficient for Fat. Calculate and interpret a 95% confidence interval for the estimate of the slope for Fat. Is there evidence of a significant association between Fat and Rating?

```r
# fat.lm <-
```

Fit a simple linear regression model for predicting Rating taking into account whether or not there is pepperoni on each slice (use indicator coding).

```r
# pepperoni.lm <-
```

(9) Interpret the coefficient for Pepperoni. Is there evidence of a significant association between Pepperoni and Rating? What percent of the variation in Rating can be explained by Pepperoni?

(10) Is it accurate to say that the relationship between Calories and Rating is causal. Why or why not?

Key idea: Because fat is a potentially confounding variable we can either 1) adjust the rating values by fat OR 2) adjust the calories values for fat. The key idea is that in an observational study adjusted associations may be different than the unadjusted association (coefficients may change). Recall that an advantage of a balanced factorial designed experiment is that the adjusted and unadjusted associations are the same (coefficients do not change).

Multiple Linear Regression Models

Consider the two-variable linear regression model (indicator coding).

$$\mathrm{y}_i = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \epsilon_i \quad \epsilon_i \sim \mathrm{Normal}(0, \sigma^2)$$

- Interpret β_1 in the model.
- Interpret β_2 in the model.

Fit the two-variable (multiple) linear regression model with Fat and Calories (indicator coding).

```r
# two.var <-lm(
```

(11) What proportion of the variation in Rating can be explained by Calories after adjusting for Fat? Is this more than without adjusting for Fat?

(12) Let us assume the validity conditions for the theory-based test are satisfied (they really are not, we should simulate!), what conclusions can we make about the *model*? Be sure to provide evidence to support your conclusions.

```
# resid_panel(
```

(13) What does the coefficient on Calories mean in context? Is this a significant association?

(14) What does the coefficient on Fat mean in context?

(15) Calculate and interpret a 95% confidence interval for the coefficient for Calories.

```
# confint(
```

Fit the two-variable (multiple) linear regression model with Calories and Pepperoni (indicator coding).

```
# CalPepp.lm <-

# confint(CalPepp.lm)
```

(16) Interpret the coefficient for Calories and Pepperoni. Is there evidence of a significant association between the model and Rating? What percent of the variation in Rating can be explained by the model?