Lesson 16-17 Pizza.

by LTC J. K. Starling

Last compiled on 23 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.

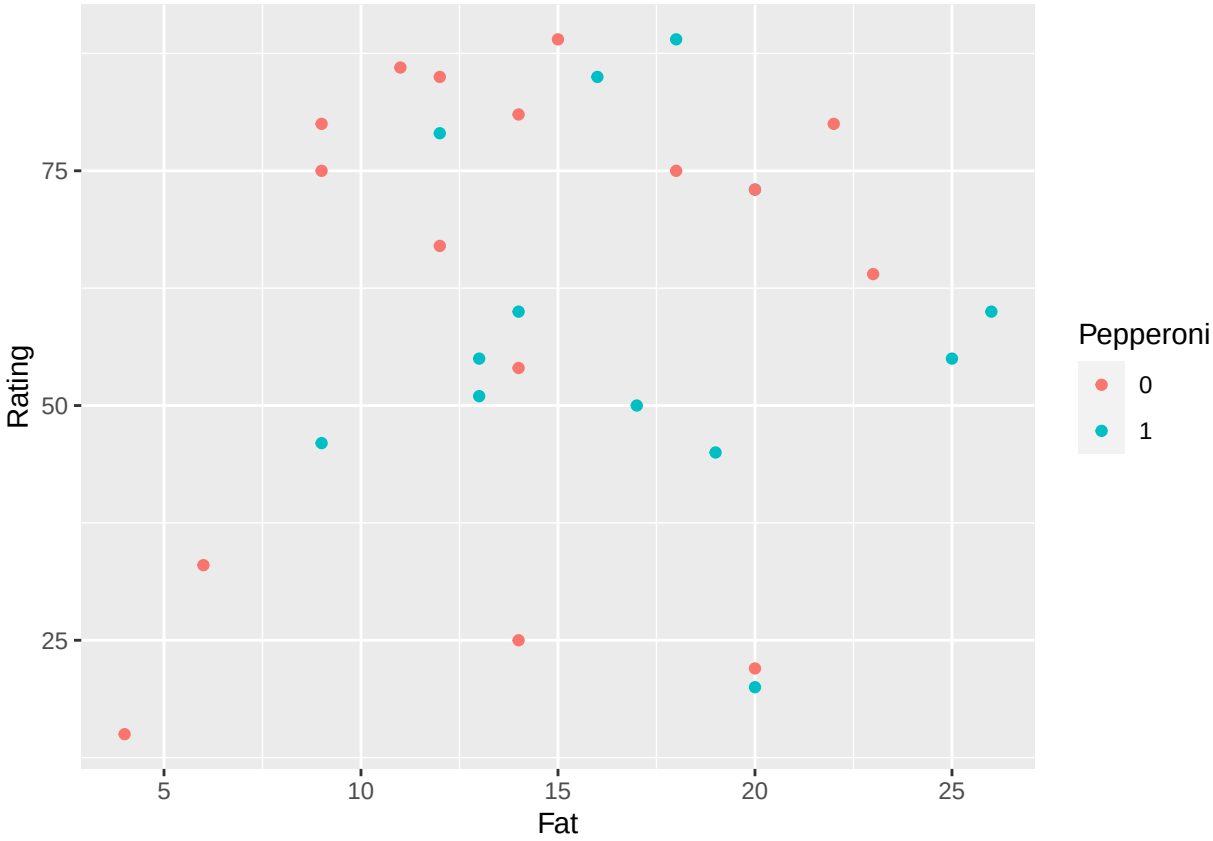
(1) Is this study an observational study or a randomized experiment? Explain.

Observational study; we are not randomizing observations to experimental groups.

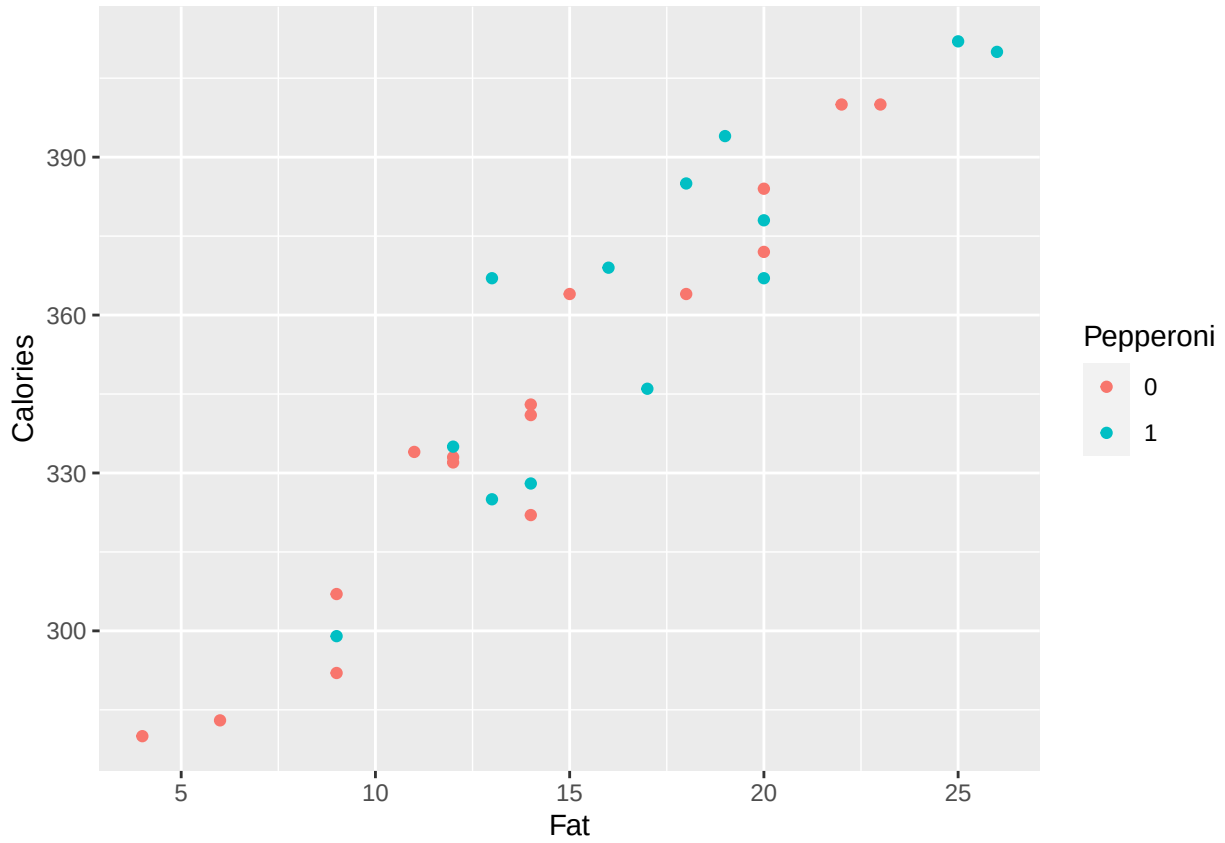
Step 3: Explore the data

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

```
pizza.dat %>% ggplot(aes(x=Fat, y=Rating, color=Pepperoni)) + geom_point()
```




```
pizza.dat %>% ggplot(aes(x=Fat, y=Calories, color=Pepperoni)) + geom_point()
```




```
# pizza.dat %>% ggplot(aes(x=Calories, y=Rating, color=Pepperoni))+geom_point()
#
# pizza.dat %>% ggplot(aes(x=Rating, y=Pepperoni, group=Pepperoni, color=Pepperoni)) + geom_boxplot() +
```

Step 4: Draw inferences beyond the data Simple Linear Regression Models

Consider the following linear regression model (indicator coding).

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

- Interpret β_1 in the model.
- Interpret β_0 in the model.

Fit the simple linear regression models (indicator coding).

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

For each extra calorie, predicted taste rating increases by 0.133 points. A p-value = 0.237 suggests that this is not a significant association at the 0.05 significance level.

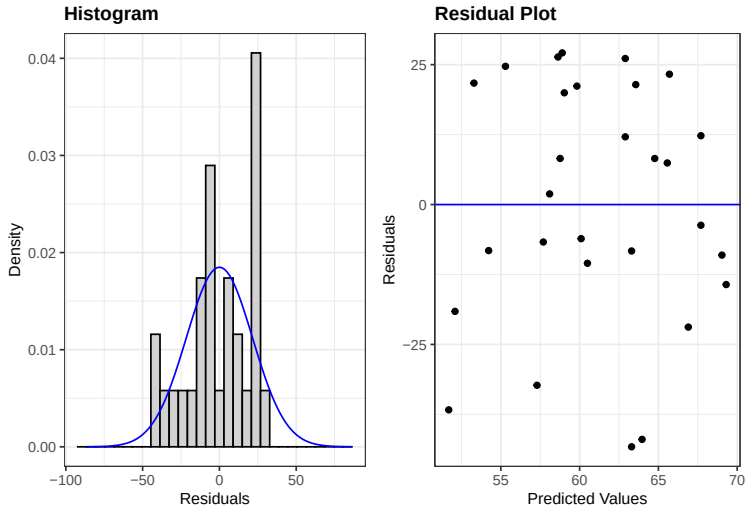
(4) Are the validity conditions for the theory-based test satisfied?

- L
- I
- N
- E

```
calories.lm <- lm(Rating ~ Calories, data=pizza.dat)
summary(calories.lm)
```

```
##
## Call:
## lm(formula = Rating ~ Calories, data = pizza.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.298 -10.496   1.905  21.171  27.105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.3392    38.8743   0.369    0.715
## Calories      0.1334     0.1103   1.210    0.237
##
## Residual standard error: 21.97 on 27 degrees of freedom
## Multiple R-squared:  0.05141,    Adjusted R-squared:  0.01627
## F-statistic: 1.463 on 1 and 27 DF,  p-value: 0.2369
```

```
resid_panel(calories.lm, plots = c('hist', 'resid'))  # Validity conditions
```



(5) Interpret the coefficient for Fat.

(6) 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?

For each extra gram of fat, predicted taste rating increases by 0.0.397 points. XXX.

```
fat.lm <- lm(Rating ~ Fat, data=pizza.dat)
summary(fat.lm)
```

```
##
## Call:
## lm(formula = Rating ~ Fat, data = pizza.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.950 -11.760  -0.139  19.223  28.033
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.0180    12.6322   4.355 0.000172 ***
## Fat           0.3966     0.7771   0.510 0.613968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.45 on 27 degrees of freedom
## Multiple R-squared:  0.009554,   Adjusted R-squared:  -0.02713
## F-statistic: 0.2604 on 1 and 27 DF,  p-value: 0.614
```

```
confint(fat.lm)
```

```
##                 2.5 %    97.5 %
## (Intercept) 29.098819 80.937211
## Fat         -1.197909  1.991066
```

(7) Interpret the coefficient for Pepperoni. Is there evidence of a significant association between Pepperoni and Rating?

(8) What percent of the variation in Rating can be explained by Pepperoni? \textcolor{blue}{0.7% }

```
pepperoni.lm <- lm(Rating ~ Pepperoni, data=pizza.dat)
summary(pepperoni.lm)
```

```
##
## Call:
## lm(formula = Rating ~ Pepperoni, data = pizza.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.750  -9.077   1.250  17.250  29.923
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.750      5.619  11.167 1.26e-11 ***
## Pepperoni1    -3.673      8.393  -0.438    0.665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.48 on 27 degrees of freedom
## Multiple R-squared:  0.007043,   Adjusted R-squared:  -0.02973
## F-statistic: 0.1915 on 1 and 27 DF,  p-value: 0.6651
```

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

No it is not causal. Fat may confound the relationship between Calories and Rating.

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).

Lesson 17

Multiple Linear Regression Models

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

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

- Interpret β_1 in the model.
- Interpret β_0 in the model.

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

```
two.var <-lm(Rating ~ Calories + Fat, data=pizza.dat)
summary(two.var)
```

```
##
## Call:
## lm(formula = Rating ~ Calories + Fat, data = pizza.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.207  -8.774   4.012  14.507  29.319
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -133.3564    81.5329  -1.636   0.1140
## Calories       0.7522     0.3222   2.335   0.0276 *
## Fat           -4.5104     2.2218  -2.030   0.0527 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.8 on 26 degrees of freedom
## Multiple R-squared:  0.1812, Adjusted R-squared:  0.1182
## F-statistic: 2.877 on 2 and 26 DF,  p-value: 0.07436
```

```
confint(two.var)
```

```
##                     2.5 %      97.5 %
## (Intercept) -300.94965130 34.23692284
## Calories       0.08990795  1.41441235
## Fat           -9.07729820  0.05650933
```

(10) 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.

(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? \textcolor{blue}{18.1%. It is more than without adjusting for Fat.}

(12) What does the coefficient on Calories mean in context? Is this a significant association? For each extra calorie, predicted taste rating increases by 0.752 after adjusting for fat. i.e. if we keep fat at a fixed value the regression equation predicts that taste rating will increase if we increase calories.

(13) Calculate and interpret a 95% confidence interval for the coefficient for Calories. \textcolor{blue}{The true coefficient for calories is between 0.09 and 1.41, with 95% certainty.(Statistically significant)}

(14) What does the coefficient on Fat mean in context? For each extra gram of fat, predicted taste rating goes down by 4.51 points after adjusting fror calories, i.e. if we keep the calories at a fixed value the regression equation predicts that taste rating will decrease with increased fat.

Multiple Linear Regression Model with an Interaction

(15) What does it mean for there to be an interaction between Calories and Fat? An interaction between calories and fat, two quantitative variables, means that there are different linear associations between taste rating and calories for pizzas with different fat content.

Key idea: The question of whether there is an interaction is a different question than whether or not adjusting for fat changes the overall association between rating and calories. (DRAW CAUSAL DIAGRAMS TO DISTINGUISH BETWEEN THESE TWO IDEAS.)

Now consider the linear regression model with the interaction term.

$$\text{Rating}_i = \beta_0 + \beta_1 \text{x}_1 + \beta_2 \text{x}_2 + \beta_3 \text{x}_1 * \text{x}_2 + \epsilon_i \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

- Interpret β_3 in the model. There is no unique effect of x_1 or x_2 on the response. WLOG, let's say we are talking about x_1- the relationship between x_1 and the response is different for every x_2 (there is no unique effect of x_1).

β_3 is the additional effect of x_1 on the response at every x_2

- Interpret β_1 in the model. Really be careful! This is the additional change in the response given a unit change in x_1 provided $x_2 = 0$

In other words if I wanted to know the total effect of x_1 on the response I would need to take into account β_1 and β_3.

- Interpret β_0 in the model. Average response when $x_1 = 0$ and $x_2 = 0$ (in the effect model it is the grand mean!)

Fit the multiple linear regression model with an interaction between Fat and Calories (indicator coding).

(16) In words, write the null and alternative hypotheses associated with β_3 for the model that includes an interaction between Fat and calories. Null hypothesis: There is no interaction between calories and fat, alternative hypothesis: There is an interaction between calories and fat.

```
interact.lm <-lm(Rating ~ Calories * Fat, data=pizza.dat)
summary(interact.lm)
```

```
##
## Call:
## lm(formula = Rating ~ Calories * Fat, data = pizza.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.962  -9.739   5.163  12.980  31.912
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -235.56044  108.34119  -2.174   0.0394 *
## Calories        1.03275    0.37450   2.758   0.0107 *
## Fat             4.52374    6.81049   0.664   0.5126
## Calories:Fat   -0.02421    0.01729  -1.400   0.1737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.43 on 25 degrees of freedom
## Multiple R-squared:  0.2407, Adjusted R-squared:  0.1496
## F-statistic: 2.642 on 3 and 25 DF,  p-value: 0.07133
```

(17) Once again, assuming that the validity conditions for the theory-based test are satisfied, what can we conclude about the *model*?

(18) For a fixed value of Calories, does the regression equation predict Ratings to go up or down as Fat increase? For a fixed value of calories, the regression equation predicts ratings will increase as fat increases.

(19) Does the interaction term significantly improve the model? Explain No. the interaction term is not significant so we should throw this model in the trash.

(20) Let's pretend that the interaction term was significant. How can we test whether a statistically significant model with an interaction is preferable to the statistically significant two-variable model? (We're also pretending that the validity conditions are satisfied for both models).

```
anova(two.var, interact.lm )
```

```
## Analysis of Variance Table
##
## Model 1: Rating ~ Calories + Fat
## Model 2: Rating ~ Calories * Fat
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     26 11249
## 2     25 10431  1    818.16 1.9609 0.1737
```

(21) What other interactions may we want to consider?

Lesson 18

(22) Can we look at the b_1 and b_2 and deduce which factor has more of an effect on predicted taste 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 taste rating.

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.

- How about the lack of association with each other?

We've also done nothing to change associations or lack of associations. The study design did not change.

- 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).

Key idea: When explanatory variables (including interactions) have a strong linear association with other explanatory variables, this can lead to larger standard errors on the slope coefficients and a larger residual standard error for the model overall.A linear association between explanatory variables is known as collinearity. This is bad! If we were looking at the interaction model, standardizing removes the linear association between the interaction and the other explanatory variables.

Key idea: Standardizing does not reduce the collinearity between main variables (though it can reduce the VIF, a commonly used indicator for concerns for collinearity).

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

```
mean(pizza.dat$Calories); mean(pizza.dat$Fat)
```

```
## [1] 350.5517
```

```
## [1] 15.34483
```

```
pizza.cov <-pizza.dat %>%mutate_at(funs(scale(.)),.vars=vars(-c(Rating, Pepperoni)))
std.lm <- lm(Rating ~ Calories * Fat,data=pizza.cov)
summary(std.lm)
```

```
##
## Call:
## lm(formula = Rating ~ Calories * Fat, data = pizza.cov)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.962  -9.739   5.163  12.980  31.912
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65.649      4.992  13.150 9.86e-13 ***
## Calories       24.894     12.160   2.047   0.0513 .
## Fat           -21.640     12.101  -1.788   0.0858 .
## Calories:Fat   -4.976      3.554  -1.400   0.1737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.43 on 25 degrees of freedom
## Multiple R-squared:  0.2407, Adjusted R-squared:  0.1496
## F-statistic: 2.642 on 3 and 25 DF,  p-value: 0.07133
```

(23) What does the intercept mean in context?

From the regression line we predict that the taste rating will be 61.10 for pizzas that have 350.6 calories and 15 grams of fat.

(24) What does the coefficient on Calories mean in context?

Predicted taste rating increases by 28.318 points for each standard deviation increase in calories when holding fat at its mean value OR Predicted taste rating decreases by 28.318 points for each unit increase in standardized calories when holding standardized fat = 0.

(25) Are the assumptions of this model valid for the data?

We assume that the observations are indpendent based on the study design. With so few data points, the residual plot is a little hard to gauge but it appears fairly scattershot (a few more underestimates that overestimates) this implies homoskedasticity (equal variance) and linearity assumptions are satisfied. The histogram is fairly symmetric suggesting the normality assumption is also satisfied.

Step 5: Formulate conclusions

(26) Based on your analysis so far, summarize the conclusions you would draw from this study.Be sure to address statistical significance, generalizability, and causation. Also be sure to put your comments into the context of this research study.

Step 6: Look back and ahead

(27) Suggest at least one way you would improve this study if you were to carry it, or a follow-up study, out yourself.