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.

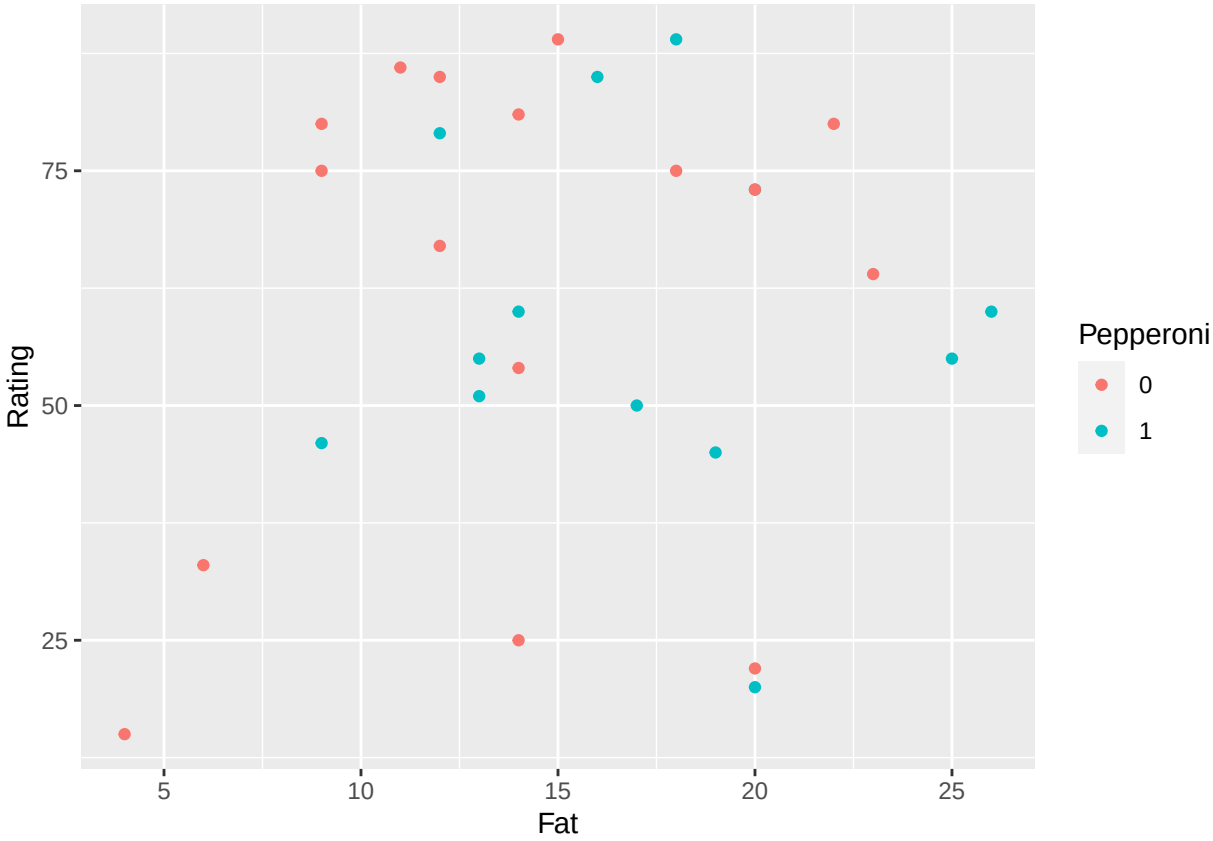
(1) Is this study an observational study or a randomized experiment? Explain.

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

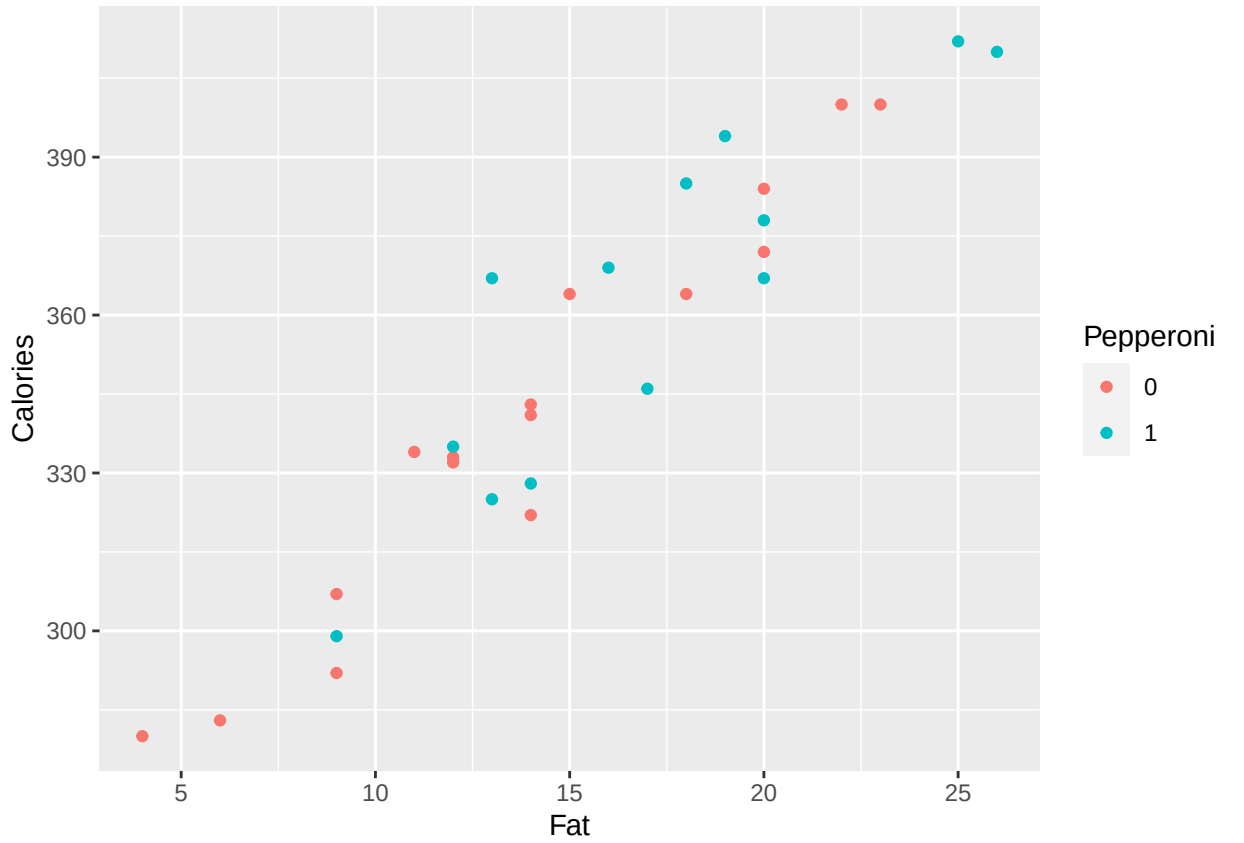
Step 3: Explore the data

(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=Lowcal)) + geom_point()
```

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'.

```
pizza.dat$Lowcal <- as.factor(ifelse(pizza.dat$Calories <=335, 'low', '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 <- pizza.dat %>% lm(Rating~Lowcal, data = .)
summary(lowcal.lm)
```

```
##
## Call:
## lm(formula = Rating ~ Lowcal, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43.50 -12.50   1.50  18.06  27.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.941      5.443  11.563 5.77e-12 ***
## Lowcallow     -4.441      8.462  -0.525    0.604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.44 on 27 degrees of freedom
## Multiple R-squared:  0.0101, Adjusted R-squared:  -0.02656
## F-statistic: 0.2755 on 1 and 27 DF,  p-value: 0.604
```

```
anova(lowcal.lm)
```

```
## Analysis of Variance Table
##
## Response: Rating
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Lowcal     1   138.7  138.75  0.2755  0.604
## Residuals 27 13599.9  503.70
```

(4) Interpret your results.

Simple Linear Regression Models

Consider the following linear regression model (indicator coding).

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

- 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 <- 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
```

```
anova(calories.lm)
```

```
## Analysis of Variance Table
##
## Response: Rating
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Calories   1   706.3  706.28  1.4632 0.2369
## Residuals 27 13032.4  482.68
```

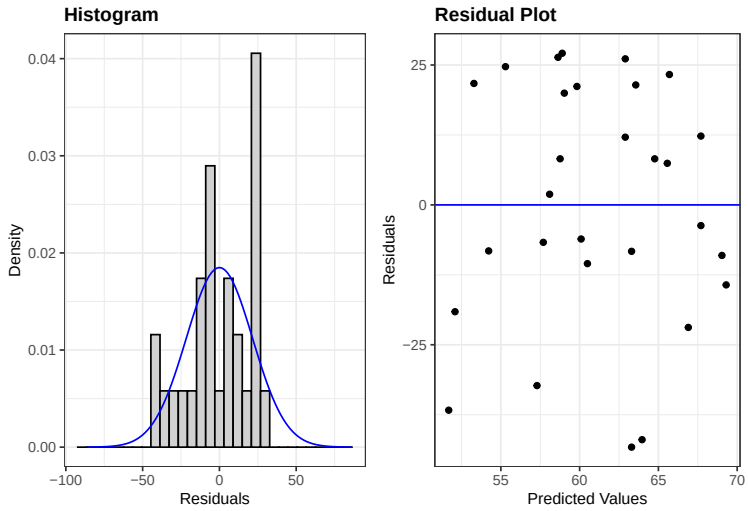
For each extra calorie, predicted taste rating increases by 0.133 points. A p-value = 0.237 suggests that this is not a significant association at the 0.05 significance level.

(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 : Linearity - the residuals vs. predicted values graph does not show any strong evidence of curvature or other patterns.
- I : Independence - the responses can be considered independent of each other.
- N : Normality - the histogram of residuals is approximately symmetric with no large outiers.
- E : Equal Variance - the residuals vs. predicted values graph shows a constant with.

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



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?

```
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
```

For each extra gram of fat, predicted taste rating increases by 0.397 points. XXX.

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

```
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) 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?

\textcolor{blue}{Pizza that has pepperoni tends to have a rating of 3.673 points lower. The R2 =0.7%.}

(10) 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).

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_2 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)
```

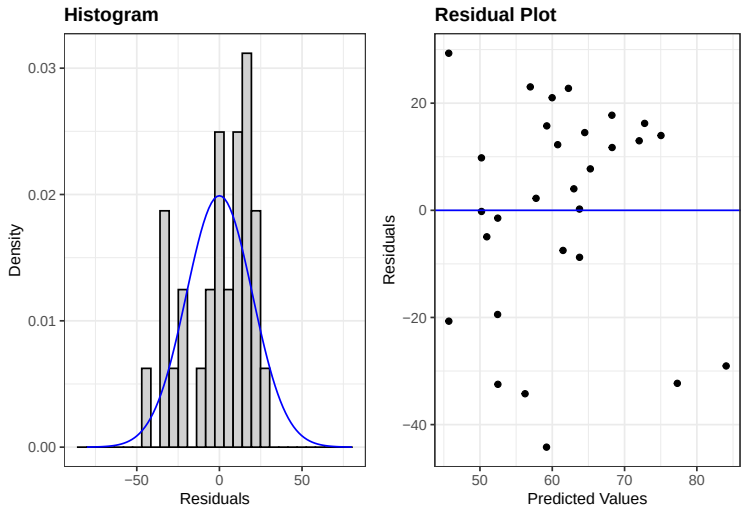
```
##
## Call:
## lm(formula = Rating ~ Calories + Fat, data = pizza.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.207  -8.774   4.012  14.507  29.319
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -133.3564    81.5329  -1.636   0.1140
## 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
```

(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) 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(two.var, plots = c('hist', 'resid'))  # Validity conditions
```



(13) 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.

(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.

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

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

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

\textcolor{blue}{The true coefficient for calories is between 0.09 and 1.41, with 95% certainty.(Statistically significant)}

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

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

```
##
## Call:
## lm(formula = Rating ~ Calories + Pepperoni, data = pizza.dat)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.829  -9.125   3.225  20.147  26.305
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.4894    39.9337   0.188    0.853
## Calories      0.1622     0.1161   1.397    0.174
## Pepperoni1   -7.2423     8.6354  -0.839    0.409
##
## Residual standard error: 22.09 on 26 degrees of freedom
## Multiple R-squared:  0.07639,    Adjusted R-squared:  0.005348
## F-statistic: 1.075 on 2 and 26 DF,  p-value: 0.3559
```

```
confint(CalPepp.lm)
```

```
##                    2.5 %     97.5 %
## (Intercept) -74.59550939 89.5743515
## Calories     -0.07642053  0.4008268
## Pepperoni1  -24.99263454 10.5079833
```

(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?

.....