

HW 5

Multiple Linear Regression Models

October 24, 2025

💡 Assignment Goals

1. Fit, interpret, and use a multiple linear regression model
2. Gain intuition for working with categorical explanatory variables
3. Fit and interpret an additive model with a categorical explanatory variable
4. Fit and interpret an interaction model with a categorical explanatory variable

Q1 (That's a-more! variables)

The [Zagat](#) guide contains restaurant ratings and reviews for many major world cities. Remy, the famous (and totally health code compliant) rat chef, has just moved to New York City with hopes of opening his own Italian restaurant. He'd like to understand what factors contribute to the average **Price** of a meal for two at an Italian restaurant in [Manhattan](#), and in particular the extent to which the Zagat customer ratings for **Food** quality (measured on a scale of 0 to 30) associate with the average price of a meal. Remy suspects that the restaurant's **Decor** ratings may confound this association, and he would also like to control for the geographic location of the restaurant, namely whether it is located **East** or west of 5th Avenue.

To answer this question, he gathers data on the Zagat ratings and meal prices for a random sample of 168 Italian restaurants in Manhattan in 2001. You can load the data into R using the following code:

```
italian <- read_csv("https://raw.githubusercontent.com/kaitlyncook/data-sets/main/italian-restaurants.csv")
```

```
Rows: 168 Columns: 7
```

```
-- Column specification -----
```

```
Delimiter: ","
```

```
chr (1): Restaurant
```

```
dbl (6): Case, Price, Food, Decor, Service, East
```

```
i Use `spec()` to retrieve the full column specification for this data.
```

```
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(italian)
```

```
# A tibble: 6 x 7
  Case Restaurant Price Food Decor Service East
  <dbl> <chr>      <dbl> <dbl> <dbl> <dbl> <dbl>
1     1 Daniella Ristorante    43    22    18     20     0
2     2 Tello's Ristorante    32    20    19     19     0
3     3 Biricchino            34    21    13     18     0
4     4 Bottino               41    20    20     17     0
5     5 Da Umberto            54    24    19     21     0
6     6 Le Madri              52    22    22     21     0
```

- (a) Fit a linear regression model that addresses Remy’s research question and print out the corresponding model summary table in R. Then, write down the form of the fitted model for the mean, defining any notation that you use when writing down the fitted regression equation.
- (b) Interpret each of the coefficients in the model, *excluding* the intercept, in context.
- (c) The total sum of squares capturing the overall variability in restaurant prices is $SST = 14421.5$; the model sum of squares, capturing the amount of this variability that is “explained” by the model fit in part (a), is $SSM = 9055.0$. Use this information to calculate, by hand, the mean square error for the model in part (a), showing all of your work. What does this number tell us/estimate about the population regression model?
- (d) Check and comment on the conditions for t -based inference with the model from part (a).
- (e) Is there an association between the quality of a restaurant’s food (as measured by Zagat customer ratings) and its prices, after adjusting for the restaurant’s decor and its location within Manhattan? Conduct an appropriate hypothesis test at the $\alpha = 0.05$ level and report the conclusions of your test in a sentence.

Q2 (It’s All Categorical to Me)

[Dr. Benjamin Spock](#) was a well-known pediatrician who faced trial in 1968 for his activities as a Vietnam War protester. Specifically, he was charged with conspiring to violate the [Selective Service Act](#) by encouraging young men to resist the draft. As part of his defense, his legal team claimed that women were underrepresented on the jury, with the theory being that female jurors were more sympathetic towards war protesters than male jurors were. The defense counsel claimed that the judge had a history of choosing juries on which women were systematically underrepresented. At that time, jury members in Boston were chosen from a [venire](#) (a group of 30 to 200 individuals pre-selected from the population by the judge’s clerk). By law, people were supposed to be selected for a venire at random. For Dr. Spock’s trial, the judge’s venire had 100 people and only 9 women, none of whom were selected to be on the actual jury.

Dr. Spock’s defense counsel collected data on the percentages of women in venires from this judge’s recent trials together with those of other judges in the Boston area.

```
# Looking at five rows of the defense counsel's data frame
jury[7:12, ]
```

```

      percent_women      judge
7             15 Spock's Judge
8              9 Spock's Judge
9             24 Spock's Judge
10            40             A
11            30             A
12            16             A

```

Table 1 summarizes the percentage of women on venires from $n = 44$ trials with six different judges:

Table 1

judge	n_trials	mean_percent_women
A	5	34.2
B	6	33.5
C	8	29.0
D	7	28.1
E	9	26.2
Spock's Judge	9	15.0

Venire composition for 44 Boston-area trials, separated by judge.

Suppose we want to fit a multiple linear regression model to explore how the percentage of women on a venire varies as a function of which judge forms the venire.

(a) Take Spock's judge to be the reference level for the **judge** variable. Using the data in Table 1, fit the corresponding multiple regression model and write down the fitted model for the mean. *Make sure that you define all notation that you use for the explanatory and response variables in this model; for indicator variables, make sure that you identify which level they indicate.*

(b) Sketch¹ a plot of the fitted regression equation from (a) and label your sketch with *each* of the terms (i.e., the intercept and all slopes) in the fitted model.

(c) Now take Judge C to be the reference level for the **judge** variable instead. Use the data in Table 1 to fit this multiple regression model and write down the fitted model for the mean. *Make sure that you define all notation that you use for the explanatory and response variables in this model; for indicator variables, make sure that you identify which level they indicate.*

(d) Sketch a plot of the fitted regression equation from (c) and label your sketch with *each* of the terms (i.e., the intercept and all slopes) in the fitted model.

(e) Compare and contrast the fitted regression models from parts (a) and (c). What about them is the same? What about them is different?

¹You will likely want to make your sketch by hand on a separate piece of paper and include a picture of that sketch in your Quarto document. The [Quarto help page](#) provides details on how to embed figures/images into a qmd file.

Q3 (The Worst Kind of White Noise)

A group of marine biologists suspects that greater contact with human activity increases the stress levels of marine wildlife. To test this theory, the scientists run an experiment in which they randomly assign 34 **shore crabs** to two groups: one group is exposed to 7.5 minutes of cargo ship noise, while the other is exposed to 7.5 minutes of ambient harbor noise. The response variable is the rate at which the crabs consume oxygen over these 7.5 minutes; animals that are stressed tend to increase their oxygen consumption.

The results of the experiment are found in the `CrabShip` data set in the `Stat2Data` R package.

```
library(Stat2Data)
data("CrabShip")
```

Use these data to answer the following questions:

(a) The researchers are interested in determining whether exposure to ship noise increases the oxygen consumption of shore crabs compared to ambient noise conditions; they would also like to account for differences in the size of the shore crabs under study (captured by their body mass, measured in grams). Visualize the relationship between these three variables by creating a scatterplot of the relationship between body mass and oxygen consumption, with the experimental noise condition that the crabs were assigned to indicated by different plotting colors. *Remember to write one or two sentences summarizing the main story or take-away of the visualization!*

(b) One team of scientists hypothesizes that, after accounting for differences in crab body mass, exposure to ship noise will increase crabs' oxygen consumption relative to ambient noise conditions; they believe that the extent of this increase will be the same for shore crabs of all sizes. Explain why the following model can be used to evaluate the scientists' hypothesis:

$$E[\text{Oxygen}_i | \text{Mass}_i, \text{Ship}_i] = \beta_0 + \beta_1 (\text{Mass}_i) + \beta_2 (\text{Ship}_i),$$

where Oxygen_i is the rate of oxygen consumption for the i th crab, Mass_i is its body mass, and Ship_i is an indicator of whether or not the crab was exposed to cargo ship noise.

(c) Fit the population model from part (b) to the `CrabShip` data. Then, write out the fitted regression equation, defining any notation that you use when doing so, and visualize your fitted model by adding it to your plot from (a).

(d) Interpret the estimated coefficients for crab body mass and for the noise exposure condition *in context*.

(e) Do oxygen consumption rates differ for shore crabs exposed to different types of noise (ambient vs. cargo ship), after accounting for their body mass? Write down the null and alternative hypotheses for this test in symbols. Then, conduct the hypothesis test² and report your conclusions in a sentence. Take your significance level to be $\alpha = 0.05$.

(f) Uh oh! A rival group of scientists argues that the model presented by the first group of marine biologists is actually *misspecified*: these rival scientists believe that the effect of ship noise on oxygen

²You may assume that all necessary conditions for t -based inference are satisfied; you do not need to check them.

consumption actually depends on a crab's mass. Explain why the following model can be used to evaluate this second group's hypothesis:

$$E[\text{Oxygen}_i | \text{Mass}_i, \text{Ship}_i] = \beta_0 + \beta_1 (\text{Mass}_i) + \beta_2 (\text{Ship}_i) + \beta_3 (\text{Mass}_i) (\text{Ship}_i),$$

where Oxygen_i , Mass_i , and Ship_i are defined as before.

(g) Fit the population model from part (f) to the **CrabShip** data. Then, write out the fitted regression equation, defining any notation that you use when doing so, and visualize your fitted model by adding it to your plot from (a).

(h) Calculate and interpret, in context, the estimated relationship between crab body mass and oxygen consumption in each of the two noise conditions. *Please show all of your scratch work.*

(i) Calculate and interpret, in context, the estimated difference in mean oxygen consumption for 50-gram crabs exposed to ship noise versus 50-gram crabs exposed to ambient harbor noise. *(Please show all of your scratch work.) Without doing any extra calculations, would you expect the estimated difference to be the same for 70-gram crabs as it is for 50-gram crabs? Why or why not?*

(j) Does the difference in mean oxygen consumption rates for shore crabs exposed to different types of noise (ambient vs. cargo ship) differ depending on the crabs' body mass? Write down the null and alternative hypotheses for this test in symbols. Then, conduct the hypothesis test³ and report your conclusions in a sentence. Take your significance level to be $\alpha = 0.05$.

Q4 (Interpreting Transformed Models)

Consider the following fitted regression model for the relationship between the land mass (*area*, measured in km^2) and number of unique mammal species (*species*) for 13 islands in Southeast Asia:

$$\log(\widehat{\text{species}}) = 1.625 + 0.2355 \log(\text{area}).$$

Recall that, in statistics, $\log()$ is used to denote the natural logarithm.

(a) What can we say about the species count comparing two populations of islands, where one population has an area that is twice as large ($\text{area} = 2x^*$) as the other ($\text{area} = x^*$)? *Please show all of your work and interpret your final result in a sentence. Be explicit about what it is that is changing about the species count (the mean? the median? something else?)*

(b) What can we say about the species count comparing two populations of islands, where one population has an area that is a third as large as the other? *Hint: what is the value of c here?* Please show all of your work and interpret your final result in a sentence. Be explicit about what it is that is changing about the species count.

³You may again assume that all necessary conditions for t -based inference are satisfied; you do not need to check them.

Q5 (Article Critique)

Revisit the *Journal of Medical Virology* article (Marín-Hernández, Schwartz, and Nixon 2020) “Epidemiological evidence for association between higher influenza vaccine uptake in the elderly and lower COVID-19 deaths in Italy”, which you first read as part of Homework 1.

Read the full article again, giving particular attention to the two paragraphs in which the authors discuss their simple linear regression analysis; these are the last two paragraphs at the bottom of the first column on page 1. Then, write a paragraph (at least three to five sentences) critiquing the authors’ work. Your critique might touch on the choice of model and the analysis itself; it might also touch on the clarity with which the authors interpret and write about their work. Some additional questions to get you thinking include:

- What regression model did the authors fit? Do you agree with this choice?
- Do you believe that the regression conditions necessary for t -based inference are plausible for this data/model? Why or why not? *We will talk about F -tests in a few weeks’ time. These tests require the same regression conditions as t -tests in order to be valid.*
- How do the authors interpret their fitted model? Do you agree with this interpretation? If not, what might you change and why?

References

Marín-Hernández, Daniela, Robert E Schwartz, and Douglas F Nixon. 2020. “Epidemiological Evidence for Association Between Higher Influenza Vaccine Uptake in the Elderly and Lower COVID-19 Deaths in Italy.” *Journal of Medical Virology* 93 (1): 64. <https://doi.org/10.1002/jmv.26120>.