OpenIntro online supplement

Interaction terms

Prerequisites: Sections 1.1-1.6, 3.1, 4.1-4.4, 7.1-7.4, and 8.1 from OpenIntro Statistics are the bare minimum.

Example 1 Suppose we were to conduct an experiment where we measured the effect of water and sunlight on plant growth. While each of these contributes individually to plant growth, we might wonder whether there is any interaction between them when promoting growth.

First and foremost, we would notice no amount of water is sufficient for plant growth if sunshine is completely absent, and vice versa. If you modeled growth simply as a function of sunshine plus water (for example, using a basic multiple regression model introduced in Chapter 8 of OpenIntro Statistics), you'd run into trouble at first. This section tackles this challenge through the use of interaction terms in the context of multiple regression.

Let's consider an experiment that examines the impact of Vitamin C from two sources on the growth of teeth in Guinea pigs.[1] In this experiment, each Guinea pig was randomly assigned to one of two possible levels of each variable:

- `supp` indicates a supplement type for Vitamin C, with levels `VC` for ascorbic acid and `OJ` for orange juice.
- `dose` indicates the amount of Vitamin C, which takes values of either 1 or 2 mg.

The researchers measured the length of the teeth of the Guinea pigs as the experimental outcome. The data are summarized in Figure 1, where each combination of treatments was applied to 10 Guinea pigs.

TIP: Experiments can randomize along 1 *or more* variables
In OpenIntro Statistics, we only considered experiments where the researchers randomly assigned one type of treatment. However, by randomizing more treatments, we can identify causal relationships among a set of variables. The example in this section takes a small step into the field of statistics called **experimental design**.

Our aim is to build a multiple regression model that accurately estimates the impact of each variable. We will start by building a model of the following form:

$$y = \beta_0 + \beta_{supp} x_{supp} + \beta_{dose} x_{dose} + residuals$$

The fitted model is summarized in Table 2. Notice the slightly different variable name `suppVC` in the table (rather than `supp`). This new `suppVC` variable was automatically generated by the statistical software since `supp` was a categorical variable. The new variable takes value 1 when the supplement is `VC` and 0 when the supplement is `OJ`.

Exercise 2 Write the model represented by the output shown in Table 2. The solution is in the footnote.[2]

[1] Bliss CI. 1952. The Statistics of Bioassay. *Academic Press*. We'll consider a subset of the data available (excludes dose level 0.5), which can be accessed in R via the `ToothGrowth` data set.

[2] $y = 14.83 - 2.93 x_{suppVC} + 6.37 x_{dose} + residuals$

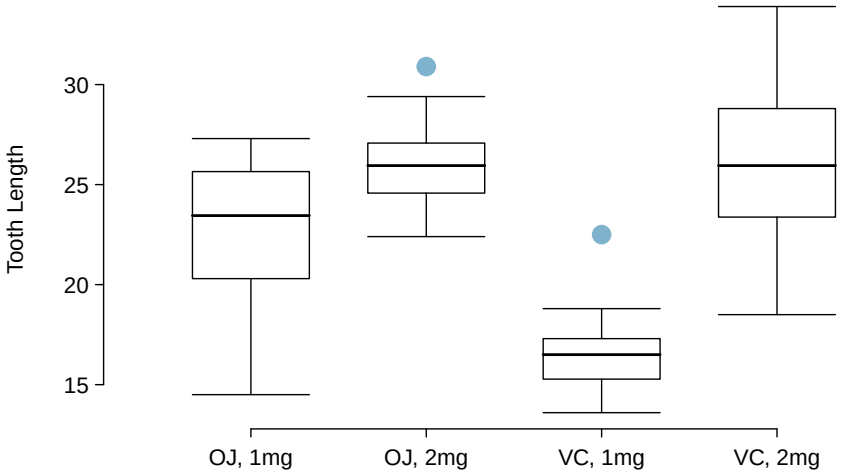


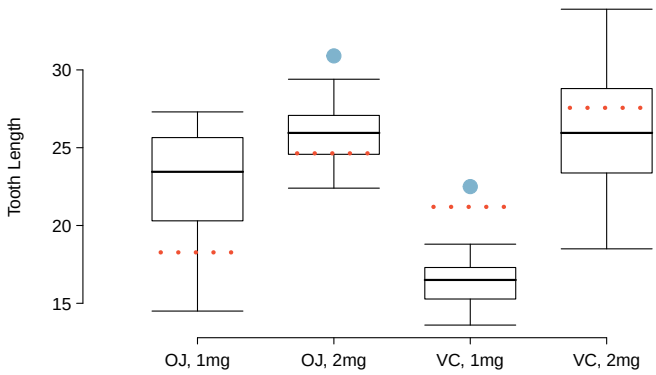
Figure 1: Side-by-side box plots summarizing the `ToothGrowth` data set. Each Guinea pig received a specific amount of Vitamin C (1mg or 2mg) and the source of that Vitamin C was either ascorbic acid (`VC`) or orange juice (`OJ`).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.8325	2.0319	7.30	0.0000
suppVC	-2.9250	1.2253	-2.39	0.0222
dose	6.3650	1.2253	5.19	0.0000

Table 2: Summary for the multiple regression model for the Guinea pig experiment.

⊙ **Exercise 3** Use the model from Exercise 2 to predict an outcome for each possible combination of variables. Calculate these means and plot them on Figure 1. The plot is provided in the footnote.[3]

[3]The means are represented by red dotted lines.



Re-examining the solution to Exercise 3, the fitted values fall far from the centers of the groups, which is a signal that the model does not fit the data very well.

The model you identified in Exercise 2 assumes that the effects of the supplement and dose are independent. However, it is also possible that the two treatments interact, i.e. the effect of one may partially depend on the value of the other. We can model this **interaction** effect using a new term:

$$y = \beta_0 + \beta_{suppVC} x_{suppVC} + \beta_{dose} x_{dose} + \beta_{suppVC:dose} x_{suppVC} x_{dose} + residuals$$

The term $\beta_{suppVC:dose} x_{suppVC} x_{dose}$ represents the interaction. The summary for this model is shown in Table 3, and the interaction term is indeed statistically significant.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.3400	2.5419	7.61	0.0000
suppVC	-11.9400	3.5948	-3.32	0.0021
dose	3.3600	1.6076	2.09	0.0437
suppVC:dose	6.0100	2.2735	2.64	0.0121

Table 3: Summary for the multiple regression model with the interaction term.

⊙ **Exercise 4** Write the model summarized by Table 3. The solution is in the footnote.[4]

⊙ **Exercise 5** Using the model equation you generated in Exercise 4, calculate the predicted value for a new observation for each group. The solution is in the footnote.[5]

[4] $y = 19.34 - 11.94 x_{suppVC} + 3.36 x_{dose} + 6.01 x_{suppVC} x_{dose} + residuals$ (could also use $x_{suppVC:dose}$ in place of $x_{suppVC} x_{dose}$).

[5] Consider the predicted value for y under each of the four possible supplement/dose scenarios:

- Orange juice and dosage of 1mg ($x_{suppVC} = 0, x_{dose} = 1$)

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_{suppVC} x_{suppVC} + \beta_{dose} x_{dose} + \beta_{suppVC:dose} x_{suppVC} x_{dose} \\ &= \beta_0 + \beta_{suppVC} \times 0 + \beta_{dose} \times 1 + \beta_{suppVC:dose} \times 0 \times 1 \\ &= \beta_0 + \beta_{dose} \sim b_0 + b_{dose} = 22.70\end{aligned}$$

- Orange juice and dosage of 2mg ($x_{suppVC} = 0, x_{dose} = 2$)

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_{suppVC} x_{suppVC} + \beta_{dose} x_{dose} + \beta_{suppVC:dose} x_{suppVC} x_{dose} \\ &= \beta_0 + \beta_{suppVC} \times 0 + \beta_{dose} \times 2 + \beta_{suppVC:dose} \times 0 \times 2 \\ &= \beta_0 + 2\beta_{dose} = b_0 + 2b_{dose} = 26.06\end{aligned}$$

- Ascorbic acide and dosage of 1mg ($x_{suppVC} = 1, x_{dose} = 1$)

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_{suppVC} x_{suppVC} + \beta_{dose} x_{dose} + \beta_{suppVC:dose} x_{suppVC} x_{dose} \\ &= \beta_0 + \beta_{suppVC} \times 1 + \beta_{dose} \times 1 + \beta_{suppVC:dose} \times 1 \times 1 \\ &= \beta_0 + \beta_{suppVC} + \beta_{dose} + \beta_{suppVC:dose} = b_0 + b_{suppVC} + b_{dose} + b_{suppVC:dose} = 16.77\end{aligned}$$

- Ascorbic acide and dosage of 2mg ($x_{suppVC} = 1, x_{dose} = 2$)

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_{suppVC} x_{suppVC} + \beta_{dose} x_{dose} + \beta_{suppVC:dose} x_{suppVC} x_{dose} \\ &= \beta_0 + \beta_{suppVC} \times 1 + \beta_{dose} \times 2 + \beta_{suppVC:dose} \times 1 \times 2 \\ &= \beta_0 + \beta_{suppVC} + 2\beta_{dose} + 2\beta_{suppVC:dose} = b_0 + b_{suppVC} + 2b_{dose} + 2b_{suppVC:dose} = 26.14\end{aligned}$$

⊙ **Exercise 6** Suppose we were to run an experiment where 24 bean plants are randomized into one of four groups:

- Each plant receives 1 teaspoon of water and 1 hour of sunlight each day.
- Each plant receives 4 tablespoons of water and 1 hour of sunlight each day.
- Each plant receives 1 teaspoon of water and 8 hours of sunlight each day.
- Each plant receives 4 tablespoons of water and 8 hours of sunlight each day.

(a) Which group do you think will have the least plant growth?

(b) The most plant growth?

(c) How confident are you in your answers?

(d) Do you think the effects of the water and sunlight on plans are independent? If so, explain why. If not, explain how you might model this relationship.