# 1 · *Explanations and Evidence*

## For practice (not to turn in)

The following scenarios will help you to hone your intuition for assessing the quality of evidence offered in support of a claim about cause and effect. For each of the following situations, identify the predictor and response variables, and decide whether the evidence can be used to support the hypothesis that $X$ causes $Y$. Discuss the issue of confounding, if relevant. You may find it helpful to draw directed graphs.

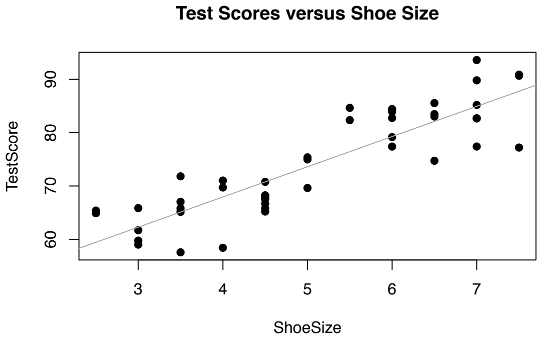
*Example 1:* A group of behavioral economists at MIT investigated a conjecture that people will perform worse on a skill-based task when they are in the presence of an observer with a financial interest in the outcome. Subjects were given time to practice playing a video game that required them to navigate an obstacle course as quickly as possible. They were then told to play the game one final time with an observer present. Subjects were randomly assigned to one of two groups. One group was told that the participant and observer would each win $10 if the participant beat a certain threshold time, and the other group was told only that the participant would win the prize if the threshold were beaten.

*Example 2:* A study looked at 155 consumers who were shopping for a new car, but had reported difficulty in securing financing to make the purchase. All of these consumers were offered free credit counseling services, and 94 accepted the offer. Of the 94 consumers who received credit counseling, 84% managed to secure a car loan within 6 months. Of the 61 who did not receive counseling, only 49% managed to get a loan.

*Example 3:* To study the effectiveness of a zinc nasal spray for reducing the duration of a common cold, researchers recruited 104 subjects who agreed to report to their lab within 24 hours of getting cold symptoms. Each subject was randomly assigned to one of three groups: one received full dosage of the zinc spray, another received a low dosage, and a third received a placebo spray. The cold symptoms lasted an average of 1.5 days for the full dosage group, 3.5 days for the low dosage group, and 10 days for the placebo group.

*Example 4:* In a recent study published in the journal *Psychosomatic Medicine*, researchers reported that people who tend to think positive thoughts catch a cold less often than those who tend to think negative thoughts. The scientists recruited over 300 initially healthy volunteers. The scientists then interviewed the subjects to gauge their emotional state, and assigned them numerical scores for positive and negative emotions. Then the researchers injected a cold virus into each subject's nose. The subjects were monitored for the development of cold-like symptoms. Subjects scoring in the bottom third for positive emotions were three times more likely to catch a cold than those scoring in the top third.

*Example 5:* You conjecture that having bigger feet makes children smarter. You go into an elementary school and take data on the shoe size and reading-test scores of 45 children. You compare test scores versus shoe size, and discover that your data looks like this:



*Potency of lime sulphur in repelling honeybees*

Load the data set `OrchardSprays` into R.[1] (1) Load the `datasets` package, using the command `library(datasets)` ; (2) Load the data set using the command `data(OrchardSprays)` . It is quite likely that the `datasets` package is loaded automatically when you launch R, so this step is probably unnecessary. You can also find the data in the file `OrchardSprays.csv`, in case you are using a different software package.

This data is from an agricultural experiment that assesses the potency of lime sulphur in repelling honeybees from trees in an orchard. Use the command `?OrchardSprays` to read a detailed description.[2]

Fit a simple group-wise model for the variable "decrease" in terms of "treatment," which is a grouping variable labeled A (highest concentration of lime sulphur) through H (lowest concentration). Summarize the overall trend in the data, using both visual evidence (such as a boxplot or a dot plot) and quantitative summaries from the model.

[1] The "exercises01-SSC325.R" script walks you through this one.

[2] "Individual cells of dry comb were filled with measured amounts of lime sulphur emulsion in sucrose solution. Seven different concentrations of lime sulphur . . . were used, as well as a solution containing no lime sulphur. The responses for the different solutions were obtained by releasing 100 bees into the chamber for two hours, and then measuring the decrease in volume of the solutions in the various cells."

## To turn in

Due Wednesday, January 23

### *(1) Confounding*

Read the article entitled "Mom's Meth Use May Affect Kids' Behavior," published on the ABC News website on March 19. 2012.

Write a few substantial paragraphs that address the following questions. Do not exceed 1 single-spaced page.[3]

1. What is the primary causal claim made by the researchers who conducted the original study?
2. Are there any issues of confounding and endogeneity that the authors of the original study would have had to consider in making this causal claim? In your estimation, does the author of the popular news article do a satisfactory job of discussing these issues?
3. In perusing the original article itself, do you get the sense that the study authors addressed these possible confounders?

### *(2) Rate of enzymatic reaction versus substrate concentration*

The data set `chymotrypsin.csv` contains data on the measured rates of the enzymatic reaction wherein chymotrypsin—an enzyme that is synthesized by the pancreas and found in your digestive system—helps break down protein molecules. For the non-biochemists like me: the rate of any enzymatic reaction obviously depends upon the underlying substrate concentration. This file contains data from a subset of a larger experiment trying to understand the nature of this relationship for chymotrypsin. This data set has 81 measurements of the reaction rate at 6 different molar concentrations of substrate ("Conc" in the .csv file). The response variable is the measured rate of the reaction. The errors in an experiment like this can be due not merely to imperfect equipment, but also to imperfections in the sample of substrate. (You might be aware that chymotrypsin is highly specific in terms of which amino acids it binds to.)

Fit a simple group-wise model of reaction rate versus concentration, treating concentration as a categorical predictor. If you want R to treat a numerical variable as a categorical variable, enclose the name of the variable in the `factor` command. For example, compare the default plots from the following two commands.

[3] This is an upper limit, not necessarily a target!

```
plot(Rate~Conc, data=chymotrypsin)
plot(Rate~factor(Conc), data=chymotrypsin)
```

Describe the overall trend in the data, including appropriate visual and numerical summaries. Do you perceive any shortcomings of a simple group-wise model here?

If you feel a bit lost with the R commands here, I highly recommend that you walk through the practice problem on orchard sprays. It has an accompanying R script that will render this problem more or less a matter of plug and play with the new variable names.

### *(3) Group-wise means*

Grouped data are sometimes written with two subscripts as $y_{ij}$, where $i$ indexes the group number, and $j$ indexes the observation within group $i$ (running from $j = 1$ to $j = N_i$, the number of observations in group $i$). Using this notation, we we would write the sample mean for the $i$th group as $\bar{y}_i$.

Prove that $\bar{y}_i$ is the unique choice of fitted value that minimizes the sum of squared residuals for group $j$'s observations. In other words, among all possible numbers $\hat{\theta}_i$, the quantity

$$\mathrm{SSE}_i = \sum_{j=1}^{N_i} (y_{ij} - \hat{\theta}_i)^2$$

is minimized when one chooses

$$\hat{\theta}_i = \bar{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \, .$$