Causal Inference Student Lab

Acknowledgements: This activity was inspired by Kahn (2005). The data first appeared in Rosner (1995). The six-step investigative process and some of the discussion questions are from Tintle et al. (2015).

Background: Today, it is common knowledge smoking has many negative health consequences. This was not always the case. The mass production and marketing of cigarettes following World War II led to a rapid increase in smoking rates, outpacing scientific and public health knowledge. Massive studies, with millions of participants, in the 1950s and 1960s led by the American Cancer Society established strong evidence connecting smoking to lung cancer. Subsequent studies, like the one this activity is based upon, sought to further refine this evidence. This activity is based on a study conducted in the 1970s. The researchers followed a cohort of children in East Boston, MA for seven years to determine, among other things, the effect of childhood smoking on lung function Tager et al. (1983, 1979).

Step 1. **Ask a research question.**

a) What is the research question of interest?

Step 2. **Design a study and collect data.** You are conducting a study of the effect of childhood smoking on lung function. You advertise your study in the local area to obtain subjects for your study, paying a small stipend and access to free medical screenings to encourage participation. You ask each participant whether or not they have smoked and measure their forced expiratory volume (FEV) in liters. FEV is a measure of lung function (see spirometry).

b) Which variable is the outcome?

c) Which variable is the treatment?

d) Is this an observational study or experiment? Explain.

e) Discuss what it means for smoking to *cause* a change in lung function. How is a causal relationship different than one of association?

f) Why is it important to show a causal relationship between smoking and lung function?

g) Confounding can occur when there are common causes of the treatment and outcome. If we identify such variables, there are various things we can do in the study design and analysis to eliminate confounding. List three potential confounding variables in this study. For each variable, state whether you think it is positively or negatively associated with the treatment and outcome. *For example, socioeconomic status could be a confounding variable. It is negatively associated with smoking and positively associated with lung function.*

For the purposes of this activity, let's say you were able to control for all other potential confounding variables other than the variables in Table 4.

Variable	Description
AGE	the age of the subject in years
FEV	forced expiratory volume (L), a common measure of lung function
HEIGHT	the height of the subject in inches
SEX	biological sex of the subject: Female (0), Male (1)
SMOKE	whether the subject had ever smoked or not: No (0), Yes (1)

Table 4: Description of variables in this study.

h) Draw a causal diagram describing the relationship between variables in Table 4. Briefly justify the inclusion/exclusion of arrows in the diagram.

i) Based on your causal diagram, which variables are potential confounders of the effect of smoking on lung function? Explain why you selected these variables.

j) Redraw your causal diagram with a box around each confounder. You will adjust for these variables in your analysis.

Step 3. **Explore the data.**

The file `FEV.csv` contains data on 654 subjects for the variables in Table 4 from a study conducted in Boston, MA in the 1970's.

k) Perform data analysis and comment on at least two interesting features of the data.

l) Calculate mean FEV for both smokers and nonsmokers and briefly discuss. For the remainder of this assignment, we'll refer to the difference in mean FEV for the two groups as the "crude association".

m) How are the confounders in your causal diagram related to smoking and lung function in the data? Discuss in terms of the size and direction of each relationship.

n) Can any of these relationships explain the crude association? Explain.

o) Estimate the effect of smoking on lung function by using an appropriate statistical method (multiple regression, stratification, matching, etc.) to adjust for potential confounders. Briefly discuss the statistical method you used and report your estimate of the effect.

Step 4. **Draw inferences.**

p) Report an appropriate 95% confidence interval for your effect estimate and briefly discuss your method.

q) Explain and assess any validity conditions for your statistical method.

Step 5. **Formulate conclusions.**

r) Is your result statistically significant? How do you know?

s) Is your result practically significant? In other words, is the reduction in FEV large enough to be meaningful? How do you know?

t) In your opinion, should we draw a cause-and-effect relationship between smoking and lung function based on your analysis?

Step 6. **Look back and ahead.**

Describe two other variables we should have considered in this analysis.