Causal Inference Student Lab
Instructor's Guide

Quick Info

Level: Introductory/intermediate undergraduate statistics courses.
Brief Description: Students investigate a research question using causal inference methods.
Topics Covered: Data Analysis, Confounding, Causal Inference, Causal Diagrams, and Multiple Regression.
Learning Goals: Students develop multivariable thinking by identifying and depicting relationships among a group of variables in a causal diagram, and by justifying their decisions. They gain practical experience with confounding variables in the context of a real analysis.
Software Required: Data analysis software such as R or Excel.
Prerequisites: Data analysis (summary statistics and basic visualizations), confounding, and multiple regression.
Time: 1 hour of class time and 1 hour of homework.
Instructor Resources: Lab, Instructor Guide, "Causal Inference in Introductory Statistics Courses".
Acknowledgements: This activity was inspired by Kahn 2005Kahn (2005). The data first appeared in Rosner 1995 Rosner (1995). The six-step investigative process and some of the discussion questions are from Tintle et al 2015 Tintle et al. (2015). The instructor guide format comes from Kuiper Kuiper (2018).

Why should you use this Causal Inference Lab in your course?
This lab fosters statistical thinking by taking a causal inference approach to investigating a research question. We designed this lab for introductory statistics courses, which typically do not cover causal inference. However, the rules of causal diagrams are simple to understand, and taking such an approach better fosters several goals of introductory courses including developing multivariable thinking, giving students experience with the entire investigative process, and fostering active learning. Also, teaching causal diagrams give instructors a useful tool for structuring discussions involving multiple variables.

What type of course is this lab designed for?
This lab is designed for a first or second undergraduate course in statistics.

When should you use this lab in your course and what are the prerequisites?
Students should have experience with using statistical software to perform data analysis and fit multiple regression models. Instructors can teach causal diagrams concurrent with the exercise or assign one of the resources listed in the paper as homework prior to the lab.

How should you conduct the lab? How much time should you expect to allocate?
We recommend having students begin the lab in class and then finish it as homework. Allowing students to work in small groups will help keep everyone moving through the activity.

Typically, students take about 1-2 hours to complete it. We recommend using 20-30 minutes in the following lesson to discuss the activity with students.

Here are some helpful hints for instructors using this lab for the first time.

- We highly recommend instructors read Kahn 2005, which proposed an activity using this same data. He discusses many interesting aspects of this study and data. We do not repeat those here, instead focusing on the causal inference aspects of the investigation.
- Allowing students to work in groups or collaborate with neighbors encourages discussion and keeps them moving through the activity.
- Instructors should consider having students read one of the two Tager papers outside of class Tager et al. (1983, 1979).
- The American Cancer Society provides a nice summary of the history of the evidence of the smoking-lung cancer link at this site: *https:// www.cancer.org/ latest-news/ the-study-that-helped-spur-the-us-stop-smoking-movement.html*

Causal Inference Student Lab
Instructor's Guide

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). The instructor guide format comes from Kuiper (2018).

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?

The research question is "what is the effect of smoking on lung function in teenagers?"

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? FEV

c) Which variable is the treatment? SMOKE

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

This is an observational study because we are not able to randomize whether a subject smoked or not.

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

In a population, smoking causes a change in lung function if there is a difference in average lung function between the entire population if they had smoked and the entire population if they had not smoked. Smoking is associated with a change in lung function if there was a difference in lung function between the smokers and nonsmokers we observed.

Some points of discussion:

- Incorrectly, students frequently think of causation as a "strong" version of correlation. Higher correlation does not imply causation.
- Also incorrectly, many students interpret causality as meaning "when one event occurs, the other event must occur."
- A correct understanding of causality involves comparing counterfactual outcomes: what would have happened if we could have observed individuals both as smokers and nonsmokers.
- Association is what you typically observe: the outcome in the smokers compared to the outcome in the nonsmokers you observed.
- For a nice visual depiction of causation versus association, see page 11 of Hernan and Robins online causal inference text *https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/*. Hernan and Robins (2018).

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

Such evidence is often required to justify significant policy interventions and change human behavior. In the absence of such evidence, it can be argued (by tobacco companies, for example) there are other causes of reduced lung function.

Some points of discussion:

- It is worth noting the tobacco industry argued the relationship between smoking and negative health effects was one of association for a long time.
- cite Cobb article here

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.*

Here are some examples:

(a) Sex. Males are more likely to smoke and have higher lung function.

(b) Physical Activity. Physical activity is negatively associated with smoking and positively associated with lung function.

(c) Diet. Eating healthy 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 5.

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 5: Description of variables in this study.

h) Draw a causal diagram describing the relationship between variables in Table 5. Briefly justify the inclusion/exclusion of arrows in the diagram.

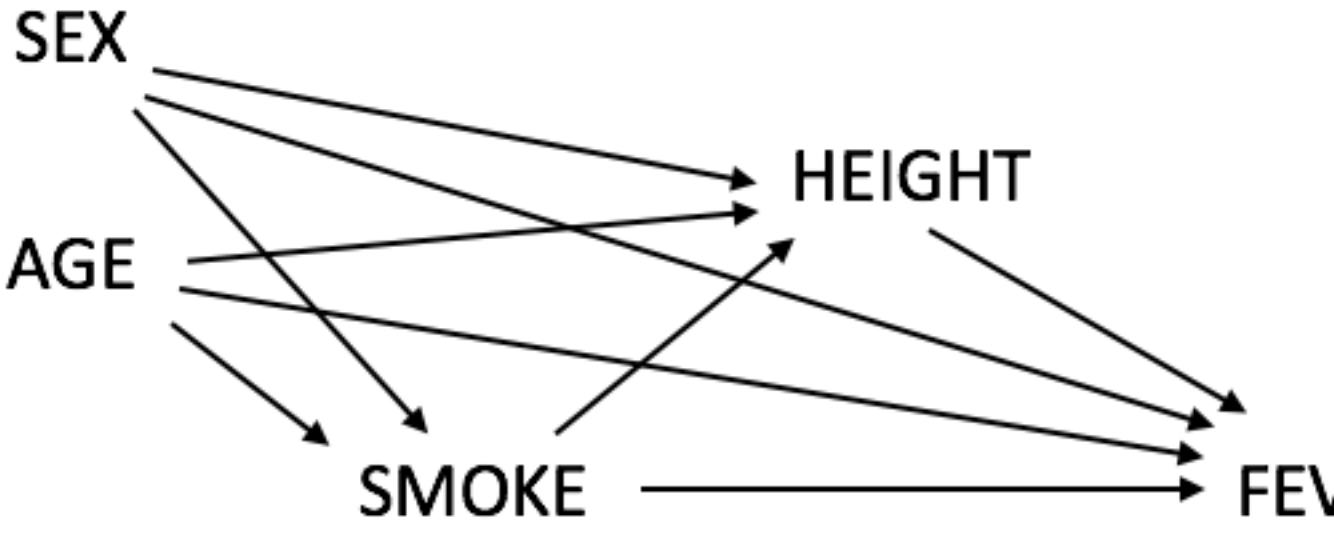

Figure 6: Causal diagram describing the causal relationships between variables in Table 5.

(a) Age. Older teenagers are more likely to smoke (more peers that smoke, greater freedom from parents, etc.), have higher lung capacity, and be taller.

(b) Sex. Men are more likely to smoke than women, be taller, and have higher FEV (even when they are the same height as women).

(c) Smoking. There is evidence childhood smoking inhibits growth, so we should not exclude an arrow between smoking and height (add citations).

(d) Height. Taller people have higher lung capacity.

Some discussion points:

- FEV increases greatly during the teenage years. Students will benefit from briefly reading about spirometry (lung function testing) when drawing their causal diagram.
- Have students critique each others' causal diagrams and discuss differences as a class.
- Variables in the diagram should be time-ordered from left to right.
- Arrows should point from left to right and there should be no cycles.
- The inclusion/exclusion of the arrow between Smoking and Height makes for nice classroom discussion.
- With the inclusion of the arrow between smoking and height, height is now on the causal pathway from smoking to FEV. For more advanced courses, you could distinguish between the direct effect of smoking on FEV and an indirect effect of smoking on FEV mediated by height.

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.

There are two confounders, AGE and SEX, of the effect of smoking on lung function. AGE is a confounder because there are two backdoor paths: Smoking – Age – FEV and Smoking – Age – Height – FEV. SEX is a confounder because there are two backdoor paths: Smoking – Sex – FEV and Smoking – Sex – Height – FEV.

Therefore, we should adjust for AGE and SEX in our statistical analysis.

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

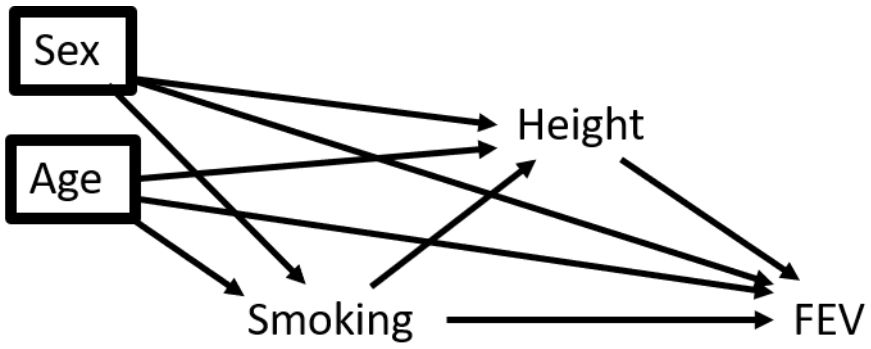

Figure 7: Causal diagram adjusting for Age and Sex.

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.

	Female		Male	
	SMOKE = 0	SMOKE = 1	SMOKE = 0	SMOKE = 1
n	279	39	310	26
AGE (years)	9.4 (2.7)	13.3 (2.2)	9.7 (2.8)	13.9 (2.5)
HEIGHT (inches)	59.6 (4.7)	64.6 (2.3)	61.5 (6.3)	68.1 (3.2)
FEV (liters)	2.4 (0.64)	3.0 (0.42)	2.7 (0.97)	3.7 (0.89)

Table 6: Mean (standard deviation) by smoking status and gender.

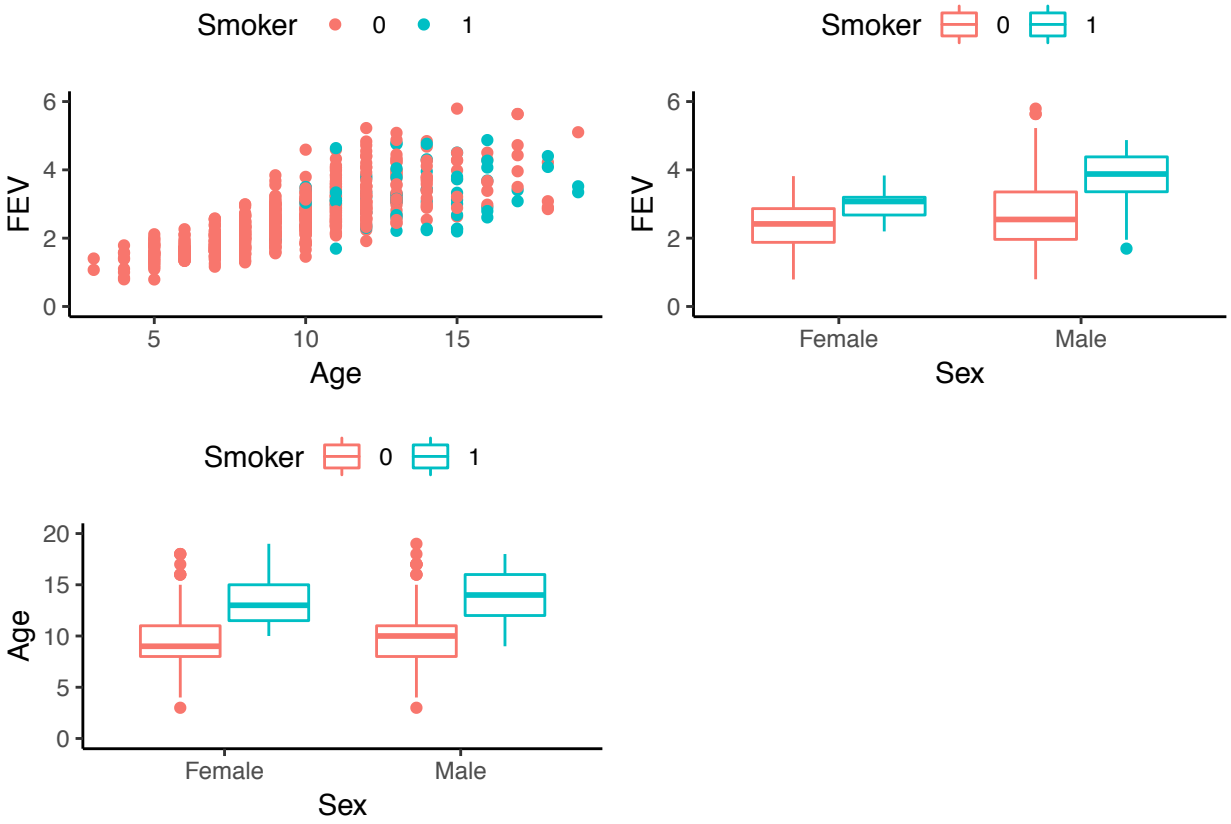

Figure 8: Summary plots of important aspects of the data set.

There is a strong positive association between Age and FEV (top left), reflecting the rapid growth in lung function of humans during teenage years. The relationship is possibly nonlinear. Kahn 2005 investigates the nonlinear relationship and is not discussed further here. Age and smoking also have a strong positive association (bottom left). Being Male is positively associated with FEV when compared across levels of smoking.

Some discussion points:

- Depending on the course, the instructor may choose to create these plots and provide them to students.
- The plot of Age vs FEV (top left) is worth discussing with the class because it provides a visual depiction of confounding by Age. Clearly, smokers (blue) are much older than nonsmokers (red). Older teenagers have higher FEV.

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".

The mean FEV is 3.3 liters for smokers and 2.6 liters for nonsmokers, showing smokers actually have a HIGHER FEV than nonsmokers in the crude analysis!

Some discussion points:

- Many students will still believe this shows a beneficial effect of smoking.
- We recommend focusing on the SMOKE, FEV, and AGE to illustrate confounding. The plot by Haggstrom (2018) is helpful for understanding how rapidly FEV increases during the teenage years.
- Students should recognize at this point that older teenagers have higher FEV and are more likely to smoke, making direct comparisons of FEV in smokers to nonsmokers not appropriate.

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.

Age has a strong positive association with smoking and FEV. Smoking is negatively associated with being male, which is not typical, and is an interesting point of discussion.

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

Yes, smokers are older than nonsmokers, and we would expect older teenagers to have higher FEV than younger teenagers even if they did smoke. Therefore, the positive association between smoking and FEV in the crude analysis is expected.

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.

Fitting a multiple regression model adjusting for Age and Gender, there is a -0.15 liter decrease in average FEV for smokers after we adjust we get the following output from R:

```
Call:
lm(formula = FEV ~ Smoke + Sex + Age, data = smoking)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.46707 -0.35426 -0.03811  0.32199  1.94943

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.237771   0.080228   2.964  0.00315 **
Smoke       -0.153974   0.077977  -1.975  0.04873 *
SexMale      0.315273   0.042710   7.382  4.8e-13 ***
Age          0.226794   0.007884  28.765  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5432 on 650 degrees of freedom
Multiple R-squared:  0.6093,Adjusted R-squared:  0.6075
F-statistic: 337.9 on 3 and 650 DF,  p-value: < 2.2e-16
```

Step 4. **Draw inferences.**

p) Report an appropriate 95% confidence interval for your effect estimate.

Model	Estimate	95% C.I.
Unadjusted	0.71	$(0.49, 0.93)$
AGE and SEX adjusted	-0.15	$(-0.31, 0.00)$

Table 7: Estimates of the change in FEV (liters) comparing smokers to nonsmokers for two models.

Step 5. **Formulate conclusions.**

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

Yes, the effect of smoking on lung function after adjusting for AGE and SEX is statistically significant with a p-value of 0.049.

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

For a 15 year old, nonsmoking male, the expected mean FEV is 3.93 liters. For a 15 year old, smoking male, the expected mean FEV is 3.78, a 4% reduction in FEV. In my opinion, this reduction at such a young age would be practically significant.

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

This is an observational study, so we should be very concerned about drawing a cause-and-effect relationship between the variables. Unmeasured confounders, such as diet or alcohol consumption, can bias the effect.

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

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

(a) Diet - smokers may be more likely to have poorer diet, and poorer diet may cause reduced lung development.

(b) Physical activity - smokers may be less likely to exercise, and less exercise may result in reduced lung function.