Introduction

Patrick Breheny

January 10

Introduction: statistical modeling

- Most statistical techniques can be thought of under the general framework of a *statistical model*
- Last semester, you focused on a specific type of model (linear regression), a fundamental building block which illustrates many of the main ideas of modeling in general
- The purpose of this course is to develop the tools and experience necessary to extend your knowledge of linear regression to a broader class of models known as *generalized linear models*

What are models?

Generally speaking, statistical models have the following components:

- An *outcome variable*, usually denoted y (also called the *response variable* or sometimes the “dependent” variable)
- A set of *explanatory variables*, usually denoted $x_1, x_2, \ldots$ (also called *covariates*, *predictor variables*, *inputs*, or sometimes “independent” variables)
- A probability distribution for the outcome
- A set of parameters (to be estimated based on the data) that quantify the precise way in which the explanatory variables affect the distribution of the outcome

What are models used for?

- Generally speaking, models are used for three main purposes:
  1. To describe and summarize a set of data
  2. To make predictions about the future
  3. To adjust for confounders when attempting to infer causal relationships
- The final of these purposes is the most interesting and the most slippery

Controlled experiments

- Suppose we are interested in knowing whether or not A causes B
- The best and most direct way is to conduct a *controlled experiment*: give A to some people, don't give A to other people, and see whether and how often B happens in the two groups
- Such an experiment is said to be controlled because we have control over who receives A and who does not

The Salk vaccine trial

- For example, consider the 1954 trial of the Salk polio vaccine conducted by the Public Health Service, which set out to answer the question of whether the vaccine prevented polio
- To answer this question, the children involved in the study were assigned *at random* to either receive the vaccine or receive a *placebo*
- Furthermore, the doctors making the diagnoses of polio did not know whether the child had received the vaccine or placebo
- The polio vaccine trial was therefore a *double-blind, randomized controlled trial*

The benefits of blinded, randomized controlled trials

- This is pretty much the best design there is
- Why?
- Because it eliminates the possibility of bias: there are no differences (either real or in the minds of the doctors or patients) between the treatment group and the control group other than the vaccine itself
- They differ in one and only one way: whether they received the vaccine or not
- Thus, any observed difference between the two groups can only be due to one of two things: the vaccine or random chance

The results of the trial

	Size of group	Polio cases per 100,000 children
Treatment	200,000	28
Control	200,000	71

The probability of seeing this big a difference by chance alone is about 1 in a billion; thus, the only plausible explanation is that the polio vaccine *causes* a reduction in the risk of polio

Observational studies

- Controlled experiments are different from *observational studies*
- In an observational study, the subject assigns themselves to the treatment/control group – the investigators just watch
- Smoking is a good example of an observational study – no one would be willing to be randomized to smoke for the rest of their lives just for the sake of a better study design, nor would it be ethical to ask it of them

Confounders

- However, there are important consequences of the fact that individuals make their own choices about smoking – it means that there are possibly many ways other than smoking in which smokers differ from nonsmokers
- Furthermore, one of these other ways may be the true cause of disease, and smoking may be irrelevant to the risk of disease
- Such factors are called *confounders*, and they represent a critically important source of potential bias present in all observational studies

Association vs. causation

- For example, smokers are more likely to be sexually promiscuous than nonsmokers and therefore more likely to contract HIV and develop AIDS; however, this does not mean that a person can stop smoking and lower their risk of AIDS
- To clarify this distinction, statisticians use the words *association* and *causation*
- In this example, smoking is *associated with* AIDS, but it does not *cause* AIDS
- In general, association is circumstantial evidence for causation – if smoking did cause AIDS, then that would explain the association
- However, it does not prove causation, as there may be other confounding factors present (in this case sexual behavior)

Using models to adjust for confounders

- Statistical models specify the way in which each explanatory variable affects the outcome, thereby isolating the effect of each variable
- Thus, they allow us to make a statement about what would happen if one variable were to change while all the others (*i.e.*, the confounders) remained the same
- Obtaining isolated effects conditional on the other explanatory variables remaining constant is said to *adjust for* (or *control for*) the effect of these confounders
- For example, "*Other things being equal*, for every additional 3g of salt you consume per day, you can expect your systolic blood pressure to rise by 1.2 mm Hg" or "After adjusting for sexual behavior, there is no association between smoking and AIDS"

Controlling for confounders

- In observational studies, identifying confounders and controlling for their effect is very important
- The more careful and well-conducted an observational study is, the more potential confounders it will adjust for, and the less plausible the explanation of confounding will become
- The devil, however, is in the details:
    - How do models implement this idea of “other things being equal”?
    - What assumptions are being made?
    - How dependent are the conclusions on these assumptions?

Hypothetical study of coffee drinking and lung cancer

- For example, consider a hypothetical study of whether or not drinking coffee increases your risk of lung cancer
- Suppose that, in reality, drinking coffee has no direct impact on lung cancer but that coffee drinkers are more likely to smoke, and smokers are more likely to develop lung cancer
- Thus, coffee drinking is associated with lung cancer, but does not cause it
- Now suppose we go out, collect a sample, and model the effect of coffee drinking on the risk of developing lung cancer

Failing to account for a confounder

- Suppose we fail to adjust for the confounding effect of smoking in our model
- In this case, our model will likely estimate a significant positive effect for coffee drinking on the risk of developing lung cancer
- This model is fine in terms of describing of the sample; after all, there is an association between coffee drinking and lung cancer
- It may also be fine in terms of prediction, if the connections between coffee drinking, smoking, and lung cancer remain constant over time

Failing to account for a confounder (cont'd)

- In terms of causality, however, the model's conclusions are faulty
- The model does not correctly predict the result of an intervention: if a man gives up drinking coffee, the model predicts that his risk of lung cancer goes down
- In reality, however, it will not change (unless he also gives up smoking)

The three purposes of modeling, revisited

- Note the progression:

  Description: The system is static
  Prediction: The system may be changing with time
  Causality: The system is being intentionally changed

- The model's conclusions are more and more dubious as we proceed down the list

Adjusting for confounders

- If, however, we also collected information on each subject's smoking exposure, we could use a regression model (in this case logistic regression) to control for the confounding effect of smoking when we estimate risks related to coffee drinking
- If we build a good model, then we will be able to correctly estimate that the effect of coffee drinking, *conditional on smoking status remaining the same*, is nonexistent
- In other words, the model can show that although there is an association between coffee drinking and lung cancer, after we adjust for the effect of smoking the association is eliminated

Complications

Building a "good" model, however, is easier said than done, as there are an endless number of considerations that may affect whether your model is good or not:

- Can we model smoking as a simple yes/no exposure, or does the risk of lung cancer go up with how heavy a smoker the subject is?
- If it does go up with smoking intensity, how? Is the effect of smoking two packs a day twice that of smoking one pack a day?
- What if a person used to smoke, but doesn't any more?
- What if coffee drinking has no effect on the risk of lung cancer among nonsmokers, but it does have an effect in combination with smoking to increase risk of lung cancer beyond smoking alone?
- And what about other confounders, such as air pollution?

Limitations of modeling

- Although modeling is very useful in terms of adjusting for confounders, it also has clear limitations
- It is not enough to "adjust for a confounder"; that adjustment must be made correctly (and how do you know whether you've made the correct adjustment or not?)
- Furthermore, you can only adjust for known confounders; it's always possible that a hidden factor is out there, biasing your conclusions
- Clearly, any conclusions we draw from our study of coffee drinking and lung cancer will be far more problematic than those drawn from the trial of the polio vaccine

The value of observational studies

- In conclusion, observational studies can never establish causation with the same certainty as controlled experiments can
- Hundreds of observational studies have shown that smoking is associated with various diseases, but none can prove causation
- However, most people would agree that, taken together, these carefully conducted observational studies make a very strong case that smoking is dangerous, and that alerting the public to this danger has saved thousands of lives
- Observational studies are clearly a very powerful and necessary tool

The importance of modeling

- However, the quality of observational studies varies tremendously, and the validity of the study depends on the details of the study design and the statistical model
- Understanding how statistical models work, what assumptions they make, how they can be used to adjust for potential biases, and what their limitations are is critical in terms of drawing valid conclusions from observational studies