Modeling for Starters

Daniel T. Kaplan

November 25, 2014

Does this sound familiar? "A parameter is to a population as a statistic is to a sample." source In introductory statistics, students encounter descriptive statistics such as the proportion, mean, variance, and correlation coefficient. These descriptions then form the object of inferential statistics, what one can validly say about the population parameter based on sample statistics.

There's something important missing here: the *purpose* behind the statistical analysis. Knowing why you're doing something is essential to deciding how to do it. A key word here is "modeling," the process of making choices about how to describe a system. A model is a "representation for a purpose." [ref to Starfield, Smith, and Blelock]. In building a model, you have to decide what features of the system are essential to your purpose and which can be omitted. Models are always approximations, but the quality of the model depends on how closely the approximation suits the purpose. George Box famously summarized the situation, "All models are wrong. Some models are useful."

The sorts of judgments that modeling makes explicit are not usually emphasized in introductory statistics. In a typical course, there are few choices to be made and they are usually related simply to the setting: Is the data categorical or quantitative? Is the description a proportion or a mean? Is one group being described, or the difference between two groups, or the differences among more than two groups?

It's understandable that things developed this way. Egon Pearson laid out many of the factors at work in his 1926 review of R.A. Fisher's "Statistical Methods for Research Workers" (*Science Progress*, **20**:733-737 available here.)

> Mr. Fisher has undertaken the very difficult task of attempting to put before research workers ... without any special mathematical training, a summary covering a great range of methods ..... The book is chiefly concerned with the best methods of handling small samples .... After conscientiously working through the examples, the student should feel able to apply the methods to exactly similar problems.... [[E. Pearson, (1926) "Review of Statistical Methods for Research Workers (R. A. Fisher)", *Science Progress*, **20**:733-734.]]

In the modern world, applying statistical methods is most reliably done with software. The R computing environment [CITATION], along with the RStudio [CITATION] interface, is now widely used in teaching, even at the introductory level. The `mosaic` package for R is designed to make teaching easier by applying the classic R modeling notation [REF: Chamberlain's book] approach to many widely used calculations and graphics. [REF: lattice] Examples in the following use `mosaic` notation with R.

Putting computation to the side for the moment, an up-to-date introduction to statistics should take into account the many other ways the statistical environment the has changed since 1926. In the era of "big data," the problems of "handling small samples" are more peripheral. Statistics is applied to much more diverse and complicated problems than laboratory experiments: the interpretation of observational data in the social sciences, medicine, commerce, and government. Whereas the experimentalist *sets* or controls the conditions of the system she studies, the analyst of observational data needs to account for the conditions **such as they may be**. At a start this requires identifying possible confounders, choosing which factors might be important and which not. In a word: modeling.

The descriptive statistics in the conventional course — means, proportions, correlations, etc. — are in fact models: representations that highlight some aspects of the data while ignoring others. Identifying these statistics as models is not a mere shift in terminology. When they are seen as models, new possibilities arise. Proportion, mean, correlation, etc. form a very limited set of models. There are many other choices. And why model the data? Instead, use the data to inform models about the real-world system under study.

To illustrate, consider the link between overweight and type-II diabetes. Originally named "adult-onset diabetes," it is now considered an "epidemic" among children. (See, for instance, this news report.)

One of the efforts to study overweight and diabetes is the *National Health and Nutrition Examination Study* (NHANES) which collects data on human body shape, disease, and mortality. The motivation for NHANES is not just an abstract quest for knowledge, it is to guide people to make helpful changes in their diet.

Causation is important here; the interest is in how diet shapes health, not in incidental correlations. Being an observational study, with no experimental intervention, NHANES is not ideally suited to drawing causal inferences. But experiment is not feasible. Data such as NHANES may be the best we have to guide important decisions.

NHANES collects each participant's Body Mass Index (BMI): weight divided by height squared (in kg/m^2). In terms of BMI, "normal" weight is defined as $20 \leq BMI \leq 25$. Overweight is $25 < BMI \leq 30$. BMI higher than 30 is considered obese.

To show the form of the NHANES data, here is a small excerpt of a handful of variables:

```
##      sex age  bmi diabetic waist weight
## 1   male  41 25.4        0 0.959   88.6
## 2   male  58 26.1        0 1.030   92.0
## 3 female  26 30.0        1 0.968   88.1
## 4 female  39 24.5        0 0.780   66.5
## 5   male  80 24.6        1 1.012   65.9
## 6 female  46 28.8        0 0.940   81.2
## 7 female  81 27.4        0 0.950   66.0
## 8 female  54 25.3        1 0.952   60.1
## 9   male  70 29.6        0 1.125  104.9
```

An introductory student taking on the problem of describing the relationship between diabetes and BMI would consider whether the average BMI different between diabetics and non-diabetics. In the modeling notation used in R/mosaic, this calculation would be written

```
mean( bmi ~ diabetic, data=NHANES )
```

```
##    0    1
## 24.5 30.9
```

Read this as "find the mean of `BMI` broken down according to the groups in `diabetic`. The relationship here is `BMI ~ diabetic`. `mean()` indicates the kind of model, `data=` specifies the source of the data.

Likely, you don't think of this sort of groupwise mean as a model. But it is. Statistical models are made by specifying three things:

1. The form of mathematical representation, e.g. the mean
2. The relationship among variables to be considered, here `bmi ~ diabetic`
3. The data used to fit the model, here `NHANES`

Obviously, means are not the only mathematical representation available. Using the mosaic/R notation, here are a few others.

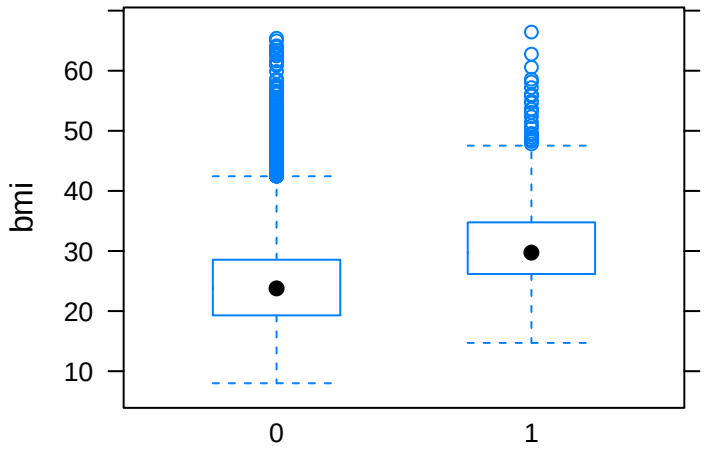
- `t.test( bmi ~ diabetic, data=NHANES )`
- `bwplot( bmi ~ diabetic, data=NHANES )`
- `lm( bmi ~ diabetic, data=NHANES )`

Each of these models has a printed form that reveals somewhat different aspects of the relationship between `bmi` and `diabetic`, and therefore can be used for a different purpose.

- The t-test reveals whether there's reason to believe that `bmi` differs between diabetics as a group and non-diabetics as a group. The purpose served by the t-test representation is to answer a question like, "Is there something here?"

```
##
##  Welch Two Sample t-test
##
## data:  bmi by diabetic
## t = -33.6, df = 1528, p-value <
## 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.80 -6.05
## sample estimates:
## mean in group 0 mean in group 1
##            24.5            30.9
```

- The box-and-whiskers plot shows how the distribution across people differs between diabetics and non-diabetics. The purpose served here is dual: "Are there outliers?" and "Are there some `bmi` values with only (or mostly) diabetics?"



The plot shows a lot of overlap between the groups. Only the very underweight values of `bmi` — say below 15 — give what might be called "protection" against diabetes. (Such low BMI, however, has its own problems: it's a sign of malnutrition.)

- The representation as a linear model is, with the very simple relationship `bmi ~ diabetic`, much the same as a t-test. Where representations such as `lm()` come into their own is when more complex relationships are being studied. More on that in a bit.

Seeing these means, t.tests, boxplots, etc. built on the relationship `bmi ~ diabetic` raises an important question: Why model `bmi` as a function of `diabetes` rather than the other way around: `diabetic ~ bmi`? After all, the purpose behind NHANES is to look for potential causes for disease, e.g. how BMI shapes the chances of developing diabetes.

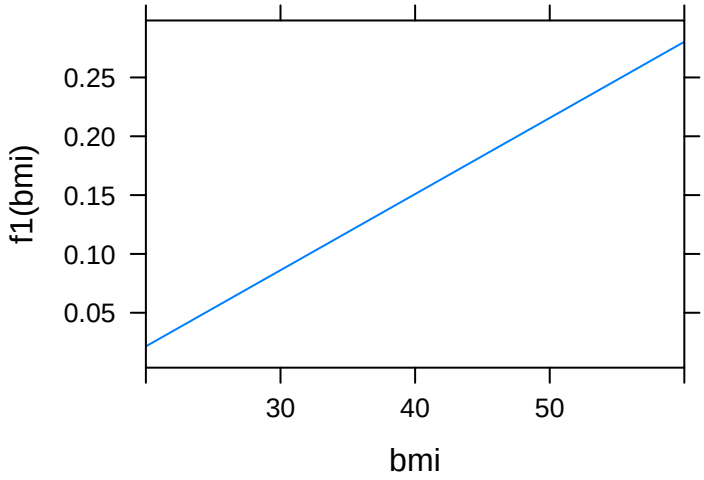
For `diabetic ~ bmi`, neither means nor t-tests nor box-and-whisker plots are meaningful. Those kinds of representations are good for looking at how a quantitative variable differs between groups. The choice to use models like the mean or t-test or box-and-whisker plot is set by the form of the data rather than the question of interest: how does BMI change the risk of diabetes.

The linear model description works with either `bmi ~ diabetic` or the more meaningful `diabetic ~ bmi`. It lets you specify the form of the relationship in a way that serves your purpose rather than merely matching the form of the data.

```
mod1 <- lm( diabetic ~ bmi, data=NHANES )
```

Although it's conventional to display such models as a regression table, there are alternative displays that are better suited for starting students. For instance, students have studied graphs of functions, slopes, etc. in high school. Why not build on this to display the model relationship in a familiar way.

```
f1 <- makeFun( mod1 )
plotFun( f1( bmi ) ~ bmi, bmi.lim=c(20,60))
```



The risk of diabetes is about 10% among those with BMI > 30, but it's half that among people of "normal" weight ($20 \leq$ BMI ≤ 30).

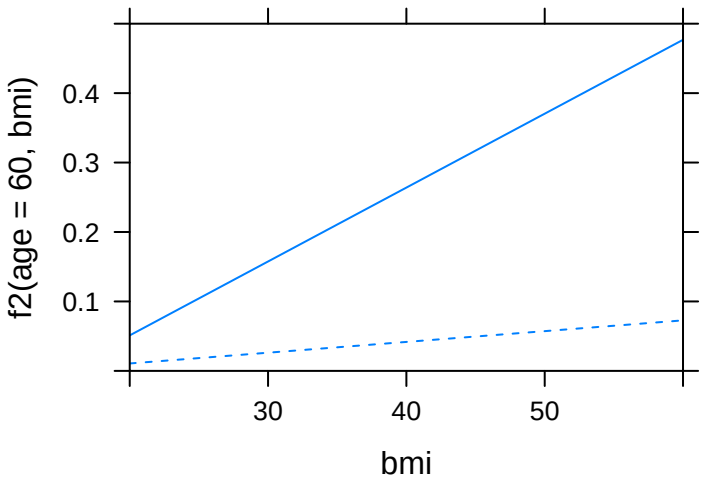
Relationships among variables are rarely as simple as `diabetic ~ bmi`. It's important to be able to model realistically complex relationships. For instance, diabetes is often associated with old age. Age is a "covariate" that might mediate the relationship between BMI and diabetes. This is easy to specify as a model: `diabetic ~ bmi * age`.

```
mod2 <- lm( diabetic ~ bmi * age, data=NHANES )
```

Any individual person is of a definite age. In displaying the relationship between `bmi` and `diabetic` for that person, it makes sense to hold `age` at that its definite value, say 25 or perhaps 60 years old.

```
f2 <- makeFun( mod2 )
plotFun( f2( age=60, bmi ) ~ bmi, bmi.lim=c(20,60),
         ylim=c(0,.5))
plotFun( f2( age=25, bmi ) ~ bmi, add=TRUE, lt=2)
```

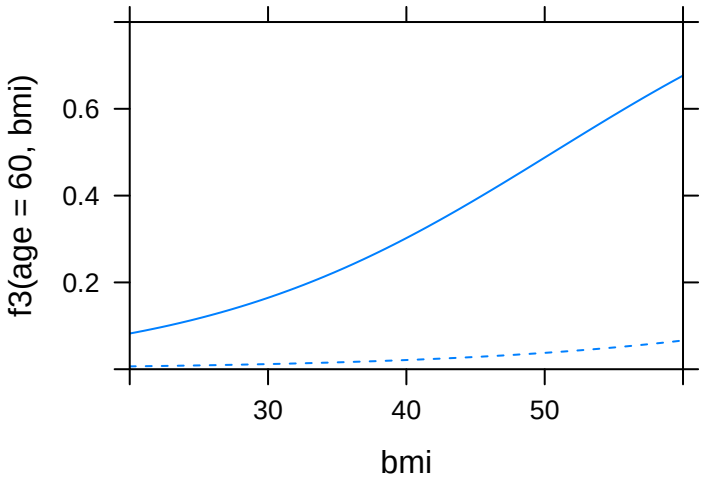
The relationship between BMI and diabetes is much stronger for 60-year olds (solid line) than for 25-year olds.



Alternative models might include other explanatory variables (smoking?). They also might have forms different from a straight line. Perhaps the risk of diabetes is not linear with age. Or, the model might respect that the risk of diabetes is always a number between 0 and 1, so perhaps use the mathematical form of the logistic function which is constrained to stay in this range.

With the modeling notation, either of these alternatives (and others) can be easily constructed. Here's a logistic-shaped model function using a quadratic function for age:

```
mod3 <- glm( diabetic ~ bmi*poly(age,2), data=NHANES,
           family="binomial" )
f3 <- makeFun( mod3 )
plotFun( f3( age=60, bmi ) ~ bmi, bmi.lim=c(20,60),
         ylim=c(0,.8))
plotFun( f3( age=25, bmi ) ~ bmi, add=TRUE, lt=2 )
```



You may ask, "Why that model and not some other? Why not include sex or other variables as a covariate? Why not split out the young and the elderly from the middle-aged?"

Our answer: Go and try it! Then you'll have to think about how to decide which of two models is better, a question which in part can be addressed through confidence intervals, etc.

Software makes it practical to **start** the study of statistics with models. There are several advantages to doing so:

- Modeling is simpler to use. One modeling framework, linear models, serves almost all the purposes that are covered by the menu of different tests taught in a conventional introductory course. (The exception is chi-squared tests. Although chi-squared is venerated in intro statistics, logistic regression provides a much more flexible framework.)
- Modeling puts tests in their place. Statisticians often complain about the "fetishization" of p-values. [CITATION FROM U CHICAGO] Models can be used to emphasize effect *size*, not just statistical significance.
- The questions addressable by modeling are much more closely tied to typical goals. Every day in the news there are stories about a claim for some relationship looking like this: "A was related to B even when C was held constant." In their own lives and work, students will benefit from thinking about what covariates might be related to the variables of direct interest and whether these covariates should be "held constant" to answer the question of interest.
- Modeling involves creative thought about how the system under study works. Modeling is not just a matter of interpreting a p-value: you can examine many factors and build a model that addresses your actual purpose. The student becomes an active participant in constructing models and comparing them, rather than merely turning a crank.
- Modeling can illuminate complicated relationships. As statistics has expanded beyond its use as a laboratory method in Fisher's data, and now is a ubiquitous tool for decision-making.

With software and comprehensible notation standing in for Fisher's "special mathematical training," students can be introduced to covariates and the process of judging whether a covariate contributes to the model in an important way. Indeed, without such capabilities, all one can do is preach caution about causation. But causation, even from observational data, is often central to the purpose of the statistical analysis.

Many instructors will feel more comfortable introducing modeling in a *second* statistics course. Many schools have a second course along the lines of "regression analysis." It can be updated as "Statistical Modeling," and be less about linear algebra and more about the human judgment that goes into constructing models.

At Macalester, our introductory class is about modeling. The course enrolls almost half of all students; it's not a specialized or "honors" class. We focus on statistical concepts [FRESH] and use R/mosaic to empower students with an accessible and expressive way to apply those concepts to data through models. Resources for helping instructors to get started with modeling are available in several forms. [REF to mosaic books and vignettes.]

The vast majority of statistics students take just the intro course, but everybody has to deal with confounding. If students don't learn this in their statistics course, where will they learn it?