Response and explanatory variables

StatPREP Class Lesson

Orientation

A variable is a quantity or characteristic that varies from one person to another, or more generally, from one *unit of observation* to another. There are two distinct *types* of variables: quantitative (numeric) and categorical (labels/words/etc.).

There's another distinction to be made about variables, which is not about the type of the variable itself but the *role* the variable will play in the statistical methods we use to describe a situation of interest and the relationships among the variables involved. This distinction is between the *response variable* and the *explanatory variables*.

Think about how we describe human relationships. For example, consider two women: Eliana and Rabia. One possible relationship: Eliana is Rabia's aunt. Another possible relationship: Rabia is Eliana's niece. As you know, these two relationships are exactly the same thing, just expressed differently. Each of the expressions involves a reference person, that is, a person with respect to whom the relationship word ("aunt", "niece") is used.

Similarly, when we describe a relationship between two variables, it's helpful to consider one variable to be the reference. We describe the pattern of the other variable *with respect to* the first variable. Now phrases like "one variable," "other variable," and "first variable" are ambiguous. It's hard to keep track of which variable is which. So, to simplify things, we identify one of the variables as the *response variable* and the other as the *explanatory variable*. Then, when describing the relationship, it's always a matter of describing the response variable *with respect to* the explanatory variable. So, if Response and Explanatory were the names of the two women, the relationship would always be stated as "Response is the _____ of Explanatory" or "As Explanatory changes by ____, Reponse changes by ____ ."

How to decide which variable should play the role of the response and which the explanatory? There's no absolute rule; generally you can describe relationships either way, keeping in mind that the statement of the relationship ("aunt", "niece") will depend on which is which. Here are some rules of thumb.

- If you are trying to *predict* the value of one variable by another, call the variable to be predicted the *response* variable. Example: Price of a rental apartment as a function of the city in which it's located.
- If you believe that one variable *causes* the other variable, call the variable being caused the *response* variable. Example: Risk of cervical cancer as explained by whether or not the person got a human papilloma virus (HPV) vaccine.
- If there's an *outcome* you're interested in, make that the *response* variable.

Example: Test scores as explained by socio-economic status.

Most statistics books have traditionally been written without making much use of the concepts of response and explanatory variables. Mathematicians, in particular, seem to like to talk about "bivariate relationships" and to use statistics, like the correlation coefficient, that are the same whichever order you put the variables in. But most two-variable statistics – for instance the slope of a regression line, or the difference between two proportions – depend on the order of the variables. As such, it's helpful to be clear indicating the particular variable you want to predict or that's being caused, or that you have a particular interest in.

And, when there are more than two variables, it's particularly helpful to distinguish between the single response variable and the multiple explanatory variables (which are sometimes called *covariates* or *confounders*.)

Activity

There are two main types of variables: quantitative and categorical. Strickly as a matter of logic, there are four possible ways that these two types can be arranged as the response and explanatory variables. It's important to know this, since the choice of an appropriate statistical technique should be shaped by the types of the response and explanatory variables.

Response	Explanatory	Statistical technique	Little App
Quantitative	Quantitative	linear regression	LA_linear_regression
Quantitative	Categorical	group-wise means	LA_t_test
Categorical	Quantitative	proportion regression	LA_proportions
Categorical	Categorical	group-wise proportions	LA_proportions

Depending on where you are in you statistics course, you might have not yet encountered one of these techniques or the other.

We're going to give you a few pairs of variables. For each pair, you are to:

1. View the data with a point plot, using the LA_point_plot Little App. This will let you readily see which type each variable is. (The choice of one variable as response and one as explanatory isn't critical here, since the reverse choice would generate the same plot but turned around the diagonal to reverse the axes.)
2. Use the point-plot Little App to visualize the relationship. Characterize it as "strong", "moderate", "weak" or "none". You don't need to give a more detailed description. We just want you to decide whether you can see a clear relationship or not.
3. In order to give a detailed description of the relationship, or sometimes

in order to detect any relationship at all, you need to calculate appropriate statistics on the data. For this, you have to pick the appropriate Little App, which often means that you need to designate one variable as the response and the other one as the explanatory variable. (You can refer to the table above.)

- Make an appropriate choice of response and explanatory variables using the rules of thumb above. Since some of these depend on the shape of your particular interest in the variables, your choice might be different from a classmate's.
- You can choose the sample size to be whatever you want, but generally a larger sample size makes it easier to see a relationship. You may also choose to *stratify* the sample.

4. Open up the appropriate Little App from step (3) and turn on whatever statistics you need to describe the relationship. (In some Little Apps, this is on by default.) Give an appropriate English-language description of the relationship. Some of the terms you might use are "upward sloping," "downward sloping", "difference of means," "difference of proportions," "no relationship."

   - Decide which one you want to designate as the response variable, and which one the explanatory variable.

5. Say whether the relationship (is any) shown by the statistics in the Little App is clearly than the relationship you discerned in step (2) from the point plot.

6. Consider causation. Sometimes causation is a matter of common sense (the rising sun causes the rooster to crow), and sometimes it can be subtle or is a matter of your beliefs about how things work in the world. No matter, it is always one of these five possibilities. The simplest to understand are these three:

   a. A causes B.
   b. B causes A.
   c. A & B are both caused in common by another factor, C.

   More subtle, and harder to understand even for professionals, are these two:

   d. There is no relationship between A and B, implying there is no causation involved or that the causation is being hidden by some other variables.
   e. A & B both cause another factor C, and the data include only some of the possible range of values of C.

   Tasks

- Thinking about the two variables you are studying, decide whether the relationship (if there is one) is in the form of (a) or (b) or (c). Keep in mind that you can't tell which one of these is from the data alone: it depends on your knowledge and opinions of how the world works.
- Does your choice of response and explanatory variables align with the causal mechanism you chose?

7. Consider prediction. Prediction may or may not relate to causation. (Example: Someone's post on a dating app may help predict what kind of person they are. But probably the kind of person they are is what causes the post to be what it is.) For the purposes of prediction, (1) the explanatory variable must be known *before* you find out the response variable. (Otherwise, why bother to predict the response variable? You know it.) And, generally, (2) the explanatory variable will be something that's relatively easy to measure, while the response is harder to measure. (Example: antigens in blood are relatively easy to measure, while past exposure to a disease is not known.)

    - Say whether your choice of response and explanatory variables make sense if your goal were to predict the response from the explanatory variable.

Some pairs of variables

These variables are found in the data available through the Little App. For simplicity, we're using only those categorical variables that have just two levels.

- `Births_2014`
  - Whether the mother is covered by the WIC program and the age of mother.
  - The age of the mother and of the father.
  - The length of gestation and the baby's weight at birth.
- `NHANES`
  - Systolic blood pressure and sex.
  - Diabetes and age
  - Weight and body mass index (BMI)
  - Income and `home_type`
- `diamonds`
  - The price of a diamond and its weight in carats.
  - The weight of a diamond and its clarity.

Version 0.1, 2019-04-15, Danny Kaplan