## Linear Regression: Lesson 1

## Response and explanatory variables

There’s an expression about how to accomplish a complex, confusing, and difficult task: *One step a at time.* The purpose of this lesson is to provide you with a step-by-step procedure for working with data. Each step can be straightforward. Some steps involve informing yourself about the data at hand. Some steps require creativity and insight about the real-world system the data represents. Some steps require application of a particular technical procedure.

Here are steps for getting started with a data frame that’s already available to you.

1. Find out what is the *unit of observation*, that is, what each row stands for.
2. Find out what are the *variables*, the name of each, and what each one stands for.
3. Choose a *response variable*.
4. Select one or more *explanatory variables* which you suspect might account for the response variable.
5. Examine *how much* of the variation in the response variable is accounted for by the explanatory variables.
6. Describe in words and numbers the *pattern of the relationship* between the explanatory variables and the response variable.

In this lesson, we’ll work with data collected by the US National Center for Health Statistics. The data frame is called Health.[[1]](#footnote-22) You’ll work with the data using a Little App about “linear regression”, which you can access at [this link](dtkaplan.shinyapps.io/LA_linear_regression/).

Open the Little App and set the data source to Health.

### Informing yourself about the data.

Steps (1) and (2) are to find out about the data frame: What is the unit of analysis and what are the variables. A standard place to get such information is the *codebook* for the data frame, which is the descriptive documentation about the data.

Go to the ***Codebook*** tab in the app. The documentation will appear.

The *unit of observation* is the kind of “thing” each row of the data is about. Read the description section of the codebook. This often contains clues about the unit of observation. Note the words “survey,” “individuals,” “interviewed.”

* What do you think is the unit of observation?

The variables are described further on in the codebook.

* Find a few variables whose meaning is obvious to you.
* Some variables have simple names like Age. Some have names that are more like a codeword, like BPSysAve. Find a few variables with such names and see if the description of the variable’s meaning helps you understand what the variable is about.

### Selecting variables

Step (3) is to select a *response variable* for you to study. The response variable is a variable of particular interest to you. Which variable is appropriate for you to choose depends on what the purpose of your work is. It’s very often the case that after completing steps (3) through (6), you’ll go back to study some other aspect of the data, and choose another response variable with which to go through the same steps.

For this lesson, let’s imagine that we’re interested in heart health and choose to focus on BPSysAve, which reports systolic blood pressure. (Systolic blood pressure is the maximum pressure created by the pumping heart, which occurs just after contraction. Blood pressure is usually reported as two numbers, say 120/80. The top number is systolic pressure.)

* Set the response variable in the Little App to BPSysAve.

Step (4) is to choose an explanatory variable. An explanatory variable is one that you think might explain something about the response variable. For instance, you might think that a high Pulse rate is what’s behind a high systolic pressure. Or perhaps you hypothesize that taller people have higher systolic pressure.

* Look through the list of variables. What are a few reasonable ones that you think might account for systolic blood pressure?
* Set the explanatory variable in the Little App to be one of the explanatory variables you’ve selected.

### How much is to be explained?

Steps (5) is about the relationship shown by the data between the response and explanatory variable you’ve selected in steps (3) and (4). This is where you make use of statistical techniques.

Step (5) is to quantify how much of the variation in the response variable can be explained or accounted for by the explanatory variable. To do this, you first have to measure how much variation there is in the response variable.

The red bars on the right end of the graph show the values for the response variable. There is one bar for each dot in the graph (although they may overlap). Note that the explanatory variable isn’t involved in the red bars, they are just about the response variable. The *variability* in the response variable is the amount of spread of the red bars. One not very reliable way to measure the variability is to look at the range of the variable, the difference between the biggest and smallest value.

There is a measuring stick in the app for measuring vertical positions and differences. Use the mouse to click on the graph and drag up or down to highlight a region. When you release, the app will show the value of the response variable at the top and bottom of the highlighted region. It will also show the difference between top and bottom.

* Use the measuring stick to find out the *difference* between the maximum and minimum value of the response variable.

The *standard deviation* us another, usually better, way to measure the spread of the red bars is with . The black I-shaped mark spans a vertical distance of one standard deviation.

* Use the measuring stick to measure the length of the standard deviation mark. Write down that number, calling it “total.”

The range and the standard deviation are different quantities. Both describe the spread of the response variable, but they do so in different ways. You could use either, but the standard deviation is a more reliable way to measure the variation, so that’s what we’ll use.

### How much is explained?

The Little App automatically shows the best-fitting straight-line description of how the response and explanatory variable are related. This is called the *regression line*. For every position on the x-axis, the regression line gives a corresponding position on the y-axis. Since the x-axis shows the explanatory variable and the y-axis shows the response variable, the straight line is a way of translating from the explanatory variable to the response variable. The actual data points are not usually exactly on the regression line, because the explanatory variable offers only a partial explanation for the reponse variable.

To measure the amount of the response variable explained by the explanatory variable, use the black bars next to the red bars. Remember the red bars showed the actual values of the response variables. The black bars are different. They show the values for the response variable that you get when you use the line to translate the expanatory variable into a value for the response variable. These values are called the “model values.” There is a blue I-shaped mark over the black bars that shows the standard deviation of the model values.

* Measure the variation of the model values with the standard deviation. You can use the measuring stick to figure out how long the blue mark is. Write down that number, calling it “explained.”

The answer to the question, “How much is explained?” is the ratio of “explained” divided by “total.” This ratio is called *R*.

* What’s the numerical value of R?

### Describe the relationship pattern.

Step (6) is to describe in words and numbers the pattern of the relationship between the response and explanatory variables. The best-fitting line shown by the Little App represents that relationship. There are many ways you might describe such a line, for instance:

* Which way the line slopes? For instance, suppose the line slopes upward from left to right. The in-words description could be simply, “The regression line has a *positive slope*.” Other possibilities are a *negative slope* or *no slope*. Another way to express this is with phrases like “Y goes *up* with X” (positive slope) or “Y goes *down* with X” (negative slope) or “Y doesn’t depend on X” (no slope). Of course, use the variable names instead of “Y” and “X”. The response variable goes in to replace “Y” and the explanatory variable replaces “X”.
* A more detailed description of the pattern describes *how much* Y goes up (or down) with X." For example, “A difference in systolic pressure of 6 units corresponds to a difference in diastolic pressure of 10 units.”
* Describe the relationship pattern between systolic blood pressure and the explanatory variable you chose.

## Activities

1. Keeping the response variable as systolic blood pressure, look through the various explanatory variables you identified earlier.

* Find the one that has the largest value of R.
* Describe the relationship pattern shown by the regression line.

Hint: As you try various explanatory variables, the red bars will stay approximately the same, particularly if the sample size is large. (A new random sample is taken each time you change a variable.) So you can focus your attention on the length of the blue I-mark showing the “explained” standard deviation.

1. Switch to the Births\_2014 data frame. Follow steps (1) through (6), choosing baby\_wt as the response variable and whatever explanatory variables make sense to you.

* Report R and the pattern description for one of the explanatory variables.

1. DRAFT: Is this how we want to standardize things? [↑](#footnote-ref-22)