How much is explained?

Helen Burn & Daniel Kaplan

6/12/2020

Orientation

The statistical work we do is built around *models*. If the word "model" reminds you of a toy plane, you're on the right track. A model is something we build in order to represent something in the real world. Usually the model is much simpler than the real-world object.

Models are useful when they serve a purpose. For instance, a toy plane can serve the purpose of teaching a child what are the main components of a plane and how they are related.

Many of the models built by statisticians are for the purpose of summarizing the relationship between variables. The modeling *framework* we are using (think modeling clay or balsa wood or a 3-d printer) involves a single *response variable* as output and one or more *explanatory variables* as input. One purpose for a model is description: if the input changes by a certain amount, how much does the output change.

Another important purpose for a model is to help us gauge *how much* of the variation in the response variable is *accounted for* by the explanatory variables. Knowing this guides conclusions about whether the explanatory variables are important in shaping the response variable. Or, looking at things the other way, knowing how much is accounted for also tells us how much remains *unaccounted for*.

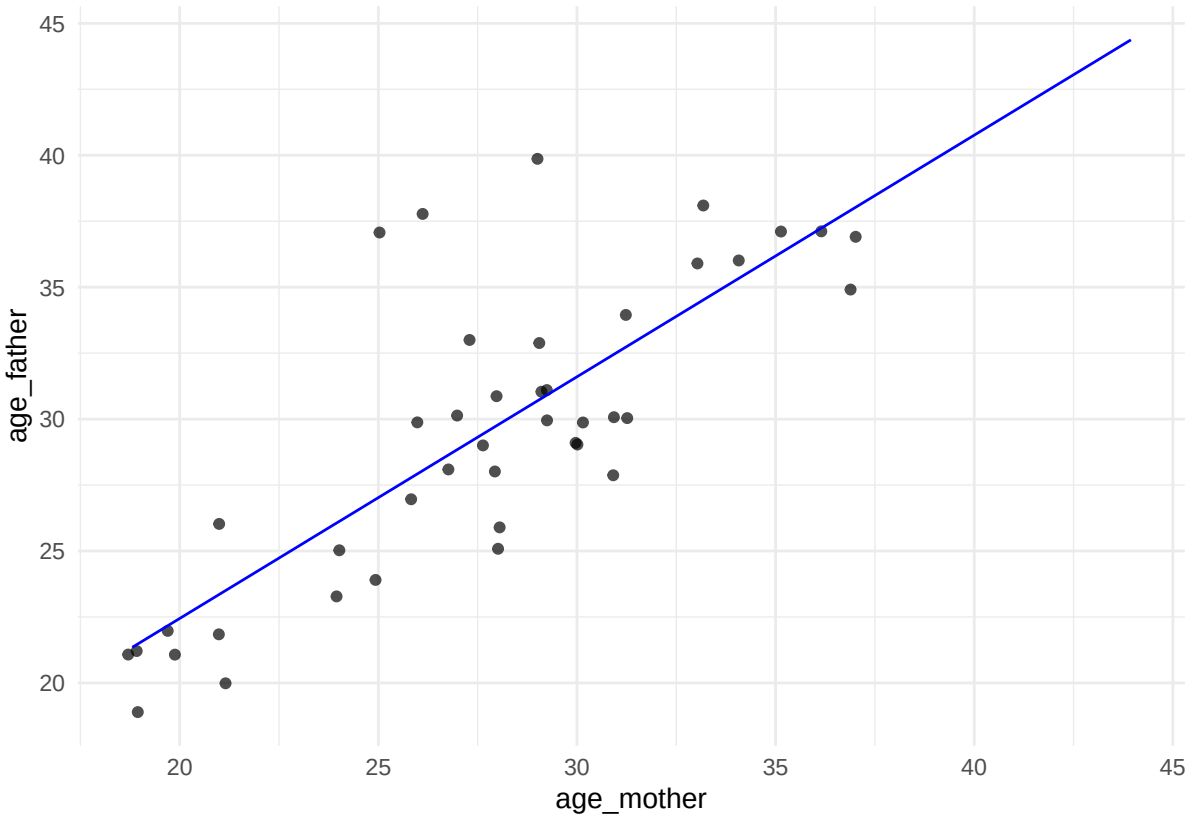
That's what this lesson is about. As you'll see, we'll work with a statistic called R with the long-winded name "coefficient of variation." (When you read about statistical models, you'll generally encounter the square of R, called "R-squared" and written R^2. There's a good reason for that, but it's not important now.)

Activity

1. The first step in finding out how much of the variation in the response variable is accounted for by the explanatory variables is, perhaps obviously, to measure how much variation there is in the response variable. Open the Little_App_Regression Little App (See footnote[1]), set Source Package to `Little Apps`, set the data set to `Natality_2014` data frame, setting the response variable to be `fagecome` the age of the father and the explanatory variable to be `mager` the age of the mother.

   The graph that will be displayed will look like this (although your random sample will be different):

[1] https://maa-statprep.shinyapps.io/Little_App_Regression/



Click on Graph in the top tool bar, and then click on the box next to Show auxiliary graph. This will display two graphs to the right, one marked as model and one marked as raw. The raw graph show the values for the response variable and an I-shaped bar. There is one dot for each value of the data (although they may overlap). Note that the explanatory variable isn't involved in the raw graph, the dots and bar are just about the response variable. The *variability* in the response variable is the amount of spread of the data. 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.

`You can use the scale on the raw graph to estimate the range.`

`*estimate the **difference** between the maximum and minimum value of the response variable. Write down`

2. Another way to quantify the spread of the model values is with the *standard deviation*. The I-shaped mark spans a vertical distance of one standard deviation.

   Estimate the length of the standard deviation mark for the raw graph.

   *Write down your measurement of the standard deviation. We'll call 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 generally use.

3. 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 response variable.

To measure the amount of the response variable accounted for by the explanatory variable, use the model graph. Remember the raw graph showed the actual values of the response variables. The model graph is different. It show the values for the response variable that you get when you use the line to translate the explanatory variable into a value for the response variable. These values are called the "model values." There is an I-shaped mark over the dots that is the standard deviation of the model values.

Measure the variation of the model values with the standard deviation using estimation.

*Write down that number, calling it "**explained**."* . . .

4. The answer to the question, "How much is accounted for?" is the ratio of "explained" divided by "total." This ratio is called *R*.

   What's the numerical value of R when using the standard deviation to measure variation? . . .

   For the graph shown above, R is about 0.6. As you can see, there is a pretty strong relationship between mother's age and father's age. You probably know the sociology of this: people tend to partner with someone of a similar age. It's not quite right to say that mother's age *causes* father's age. That's why we say that R is a measure of how much of the response variable is *accounted for* by the explanatory variables.

5. Use the Little App to explore the relationship between response and explanatory variables that you choose.

   a. *Find a pair of variables that have a large R. Write the names of the variables here.* . .

   b. *Find a pair of variables that have a small R. Write the names of the variables here.* . .

6. Select three variables: a response and *two* explanatory variables. The app calls the second of the explanatory variables a *covariate*.

   c. *Find an example where including the covariate increases R. Write down the names of the response, explanatory, and covariate here.* . . .

   d. *See if you can find an example where the covariate decreases R.* (Hint: If you make a few attempts, you'll get a good idea of what's going on.) *What did you find?* . . .

7. You might have encountered a close relative of R, written r (little-r) and called the "correlation coefficient." Little-r only makes sense when there is *only one* explanatory variable, that is, no covariate.

Version 0.3, 2020-08-13