PM R 8 (2016) 1213-1217

www.pmrjournal.org

Statistically Speaking

The Value of Scatter Plots

Kristin L. Sainani, PhD

Introduction

Authors frequently report correlation or regression statistics to quantify the linear association between 2 numeric variables. These statistics always should be accompanied by scatter plots. Scatter plots visually display the relationship between 2 variables and instantly provide the reader with key information, such as the shape of the relationship, the sample size involved, and the influence of outlying data points. "Among all the forms of statistical graphics, the humble scatterplot may be considered the most versatile, polymorphic, and generally useful invention in the entire history of statistical graphics," write Michael Friendly and Daniel Denis [1]. This article gives tips for drawing scatter plots and also gives 3 examples of how scatter plots can reveal problems with correlation and regression statistics.

Tips for Drawing Scatter Plots

Lines on plots can fool the eyes, as illustrated in Figure 1. This figure shows 2 variables (x and y) that are completely unrelated; to fool your eye, I added arbitrary diagonal lines to the plot. The lines makes it appear as if there is a positive relationship between x and y (top panel) or a negative relationship (middle panel). But when you remove the lines (bottom panel), it becomes easier to see that x and y have no relationship. Commonly, when drawing scatter plots, authors will superimpose a linear regression line to help visualize the relationship between the variables. These regression lines, however, can trick your eyes in the same way—causing you to perceive stronger relationships than actually exist.

I advise authors to examine scatter plots without overlay lines initially and then to have the computer superimpose smoothing lines rather than linear regression lines. Smoothing lines, such as LOESS (or locally weighted scatter plot smoothing) curves, are readily available in most modern statistical packages. Smoothing lines are flexible and conform to local patterns in the data. If the smoothing line reveals that the relationship between variables is not a straight line, then the authors should present the plot with the smoothing line in their paper. If the smoothing line reveals a straight-line relationship between variables, then the authors can present a plot with a linear regression line. When the smoothing line or regression line is roughly horizontal, this indicates that the variables are unrelated. Smoothing lines require the user to input a degree of curvature; generally a medium amount of curvature (not too much, not too little) is recommended. For more information on smoothing lines, see the accompanying In-Depth box.

Authors also should avoid manipulating the axes in ways that visually exaggerate the effect. Examples of such manipulations include stretching or truncating the X or Y axes.

Example 1: Scatter Plots Reveal Nonlinearity

Correlation coefficients and linear regression statistics assume that the relationship between 2 variables is linear (follows a straight line). This assumption is often violated, however, and should be checked by plotting.

In an excellent example of the use of scatter plots, the authors of a cross-sectional study of 187 cognitively normal adults [2] revealed a strong quadratic relationship between brain white matter volume and age. I used an online digitizer (http://arohatgi.info/WebPlotDigitizer/app/) to extract the values of the data points from their graph so that I could recreate the plot and reanalyze their data.

Had the authors simply calculated a correlation coefficient (or linear regression coefficient) without adequately plotting their data, they would have concluded erroneously that white matter volume and age are unrelated, because there is no linear correlation between the 2 variables ($r = .03$, $P = .70$). Figure 2 shows plots of their data. There is a strong quadratic relationship, with white matter volume increasing in volume until about 50 years of age, and then decreasing thereafter (top panel). When the authors added an

http://dx.doi.org/10.1016/j.pmrj.2016.10.018

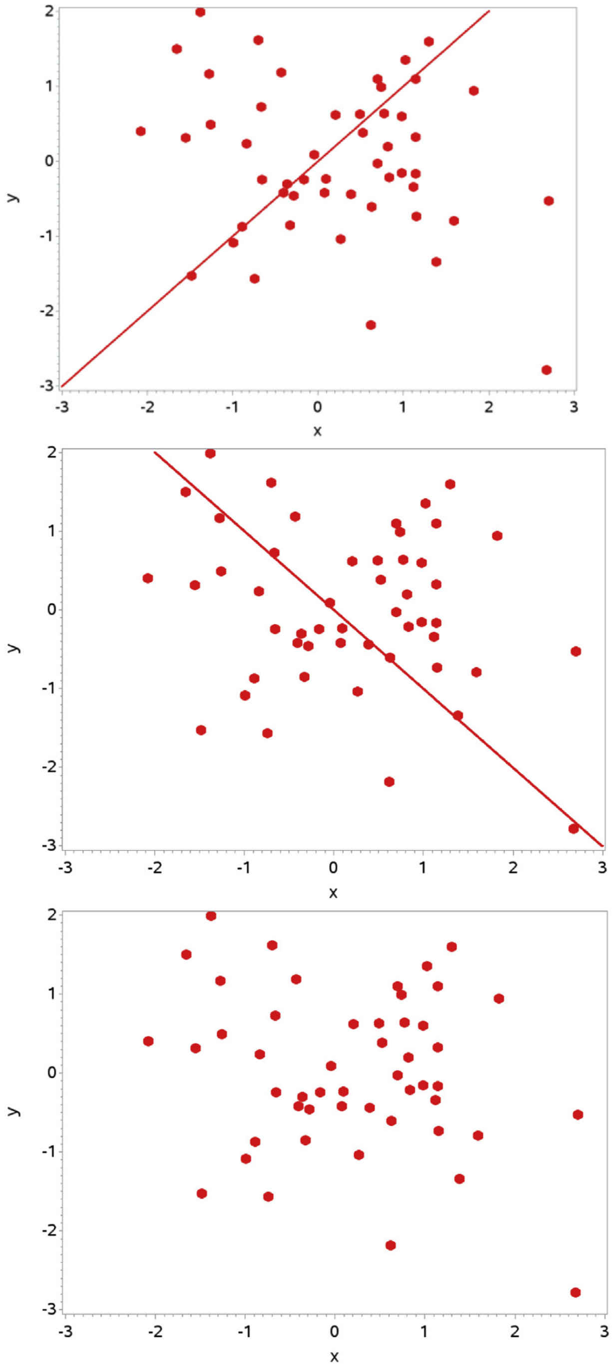

Figure 1. Lines superimposed on a scatter plot can be misleading. A +90° line has been added to the scatter plot in the top panel, which causes your eye to perceive a positive relationship. A −90°line has been added to the scatter plot in the middle panel, which causes your eye to perceive a negative relationship. There is no relationship, however, between x and y, as shown in the bottom panel.

age-squared term to their regression model, the age-squared term was highly significant ($P < .0001$), thus confirming the quadratic relationship between age and white matter volume. Because in this study the white matter volume did not strictly increase or decrease with age, a straight line fit to the data misses the parabolic relationship between the variables (bottom panel).

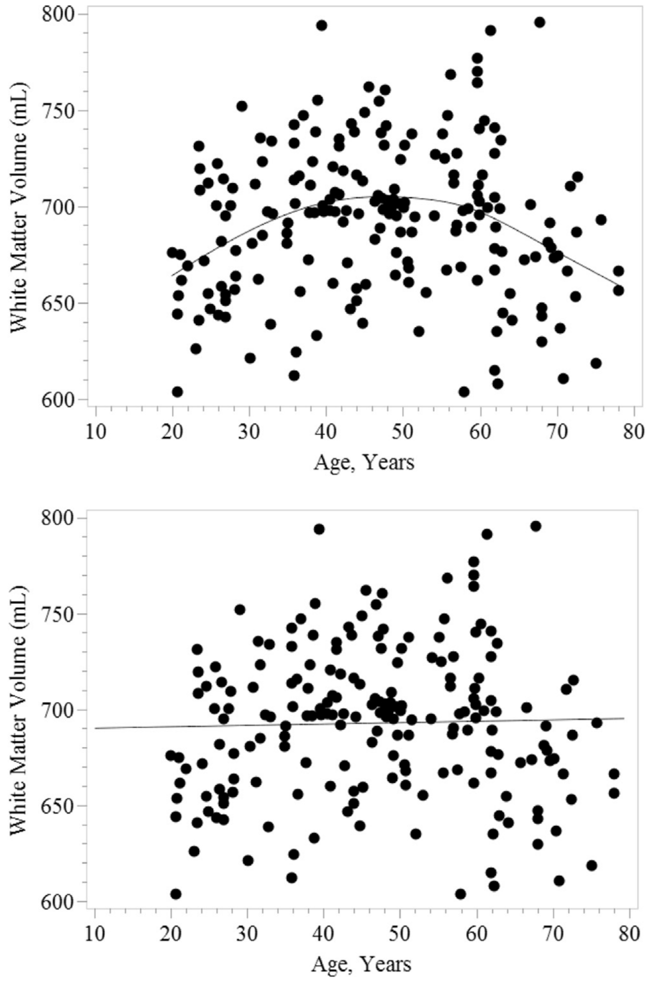

Figure 2. Example 1: Data were extracted from a cross-sectional study of 187 cognitively normal adults [2]. When plotting the data with a smoothing line, a strong quadratic relationship is revealed ($P < .001$). The bottom panel shows that there is no linear correlation between the variables (r = .03, $P = .70$).

Example 2: Scatter Plots Reveal Small Sample Sizes

Unlike box plots or bar charts, scatter plots show each individual data point, which gives readers more insight. Scatter plots visually remind readers of sample size limitations and reveal instances when a purported correlation rests on just a handful of observations.

In a 2014 study [3], 13 acne-prone young men consumed varying amounts of cocoa at baseline; a week later, a dermatologist measured their facial acne. The authors concluded that there was a "medium-strength positive Pearson correlation coefficient between the amount of chocolate each subject consumed and the number of lesions each subject developed on day 7 (r = .314)." The authors failed to mention that this correlation was not statistically significant. Commendably, they did provide a scatter plot of their data—from which I was able to extract the individual values and replicate their plot (Figure 3).