Un-modeled Confounders: Don’t get burned by Simpson’s Paradox.

As published in Benchmarks RSS Matters, December 2012
http://web3.unt.edu/benchmarks/issues/2012/12/rss-matters

Jon Starkweather, PhD

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and **S**tatistical **S**upport

http://www.unt.edu

RSS
Research and Statistical Support

http://www.unt.edu/rss

RSS hosts a number of “Short Courses”.
A list of them is available at:
http://www.unt.edu/rss/Instructional.htm

Un-modeled Confounders: Don't get burned by Simpson's Paradox.

A long time ago, in a faraway place..., the term data mining carried with it a negative connotation. However, thanks in part to Google and other pioneers of huge data analysis, data mining has become much more acceptable. The purpose of this article is to demonstrate the necessity of applying data mining principles to the initial data analysis phase of any study utilizing quantitative data. In short, it is the responsibility of the primary investigator to thoroughly explore the collected data in order to determine if the data supports the planned statistical procedures for addressing research questions or formal hypotheses; as a former colleague often said: *know thy data*. One concern common to most empirical quantitative studies is *conditional independence*. Conditional independence refers to a situation among three (or more) random variables when the relationship between two of them is independent of values of the third. Essentially, all other influences have been controlled and only the effect of interest (the relationship between two variables) is displayed in the results. In other words, ensuring that the experimental effect is isolated from any confounds, often done with probability estimates (e.g., propensity scores, matching, etc.), or the design of the study (experimental control). This article demonstrates a simple example in which the primary effect of interest is not conditionally independent of confounds. The example(s) at the bottom detail Simpson's paradox, wherein a correlation between two variables (X & Y) is strikingly misleading unless one also recognizes the underlying groups (Z) of cases. First, however, the article makes clear the definition of conditional probability and independence.

Conditional Probability & Independence

In traditional symbolic probability terms, we say X is independent of Z if:

$$p(X) = p(X|Z) \tag{1}$$

which can be interpreted as: the probability of X is equal to the probability of X given Z. This statement should make clear that X is unrelated to Z (i.e. X is independent of Z). Likewise, we could say Y is independent of Z if:

$$p(Y) = p(Y|Z) \tag{2}$$

which can be interpreted as: the probability of Y is equal to the probability of Y given Z. Now, conditional independence refers to the conditional probability of X given Z being unrelated to the conditional probability of Y given Z. Stated another way, X and Y are conditionally independent given Z, if:

$$p(X \cap Y|Z) = p(X|Z)p(Y|Z) \tag{3}$$

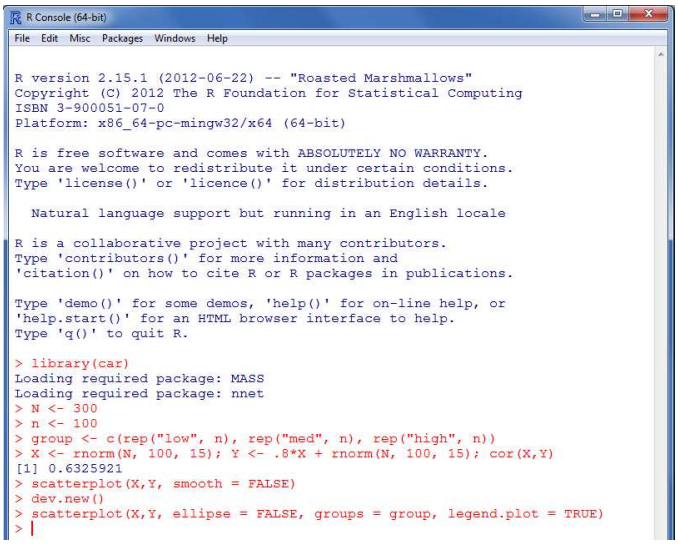
In other words, X and Y are conditionally independent given Z if they are independent in their conditional probability distributions given Z. Knowing the values of Z does not inform X or Y.

The Examples

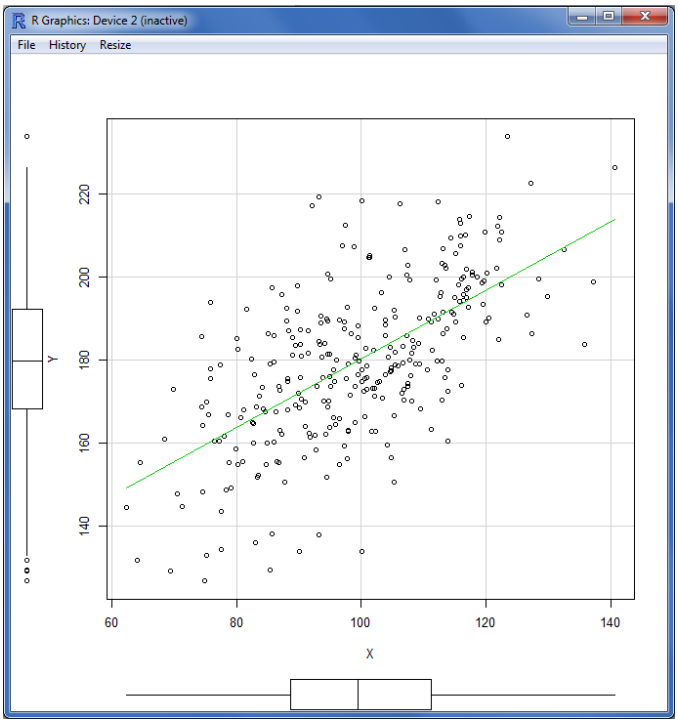
The following examples utilize the **R** statistical programming environment. Those unfamiliar with **R** can learn more at the Research and Statistical Support Introduction to **R** Short Course. The examples also utilize the function 'scatterplot' from the package car (which requires loading it and its dependencies

[package MASS & package nnet]). An example of the script used to create the images below can be found here.

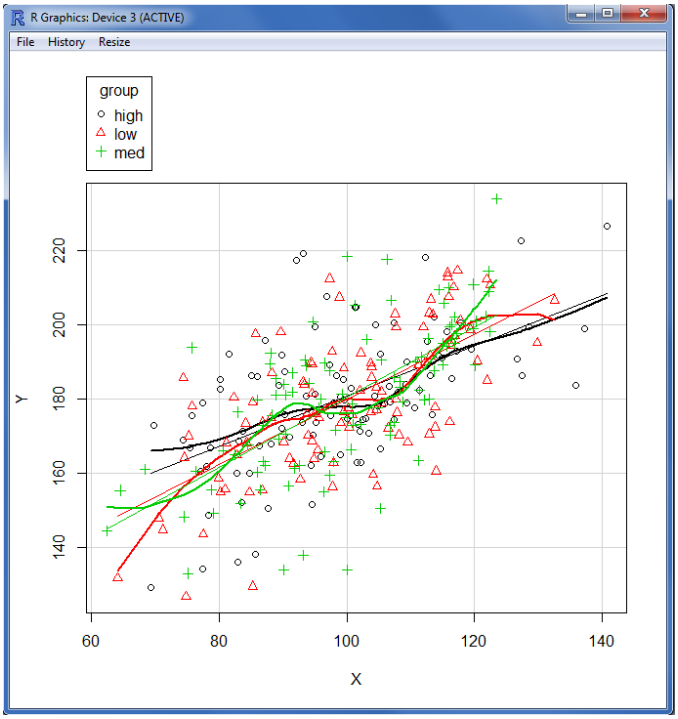
The first example illustrates conditional independence:



As can be seen above, the correlation between X and Y is positive ($r = 0.63$). We see (below) a fairly standard scatterplot generated with X and Y, without respect to the grouping variable Z. The green line represents the ordinary least squares (OLS) regression line.

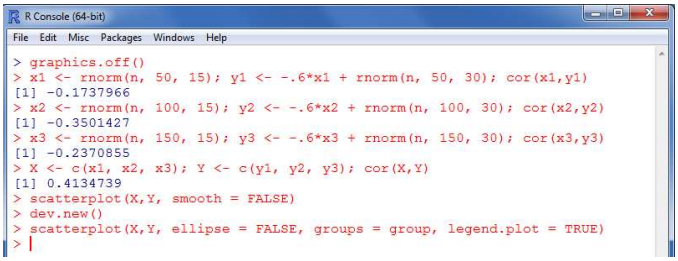


The next scatterplot shows the relationship between X and Y, as well as the groups of Z (low, medium, & high); with each of the three groups designated by different symbols and colors.

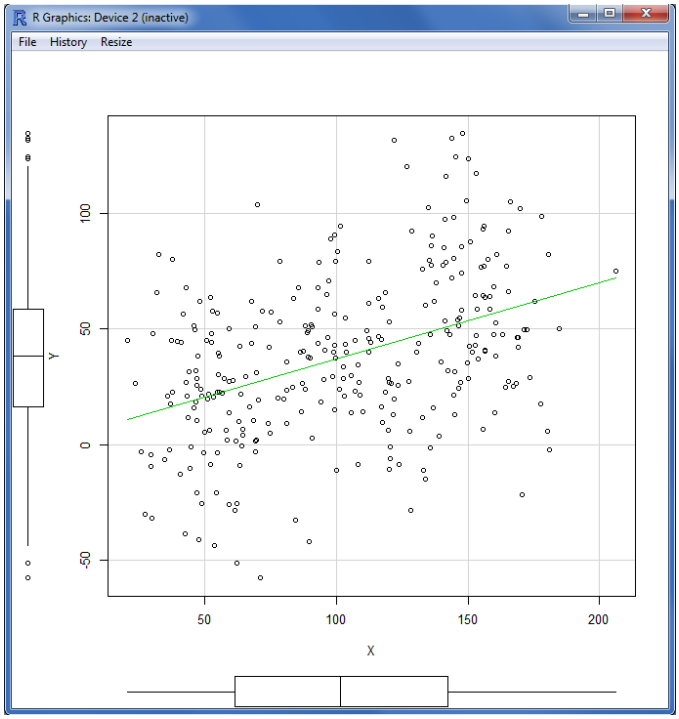


The important thing to notice in the plot above is that the relationship between X and Y is not affected by the groups (i.e. each of the groups displays essentially the same positive relationship between X and Y).

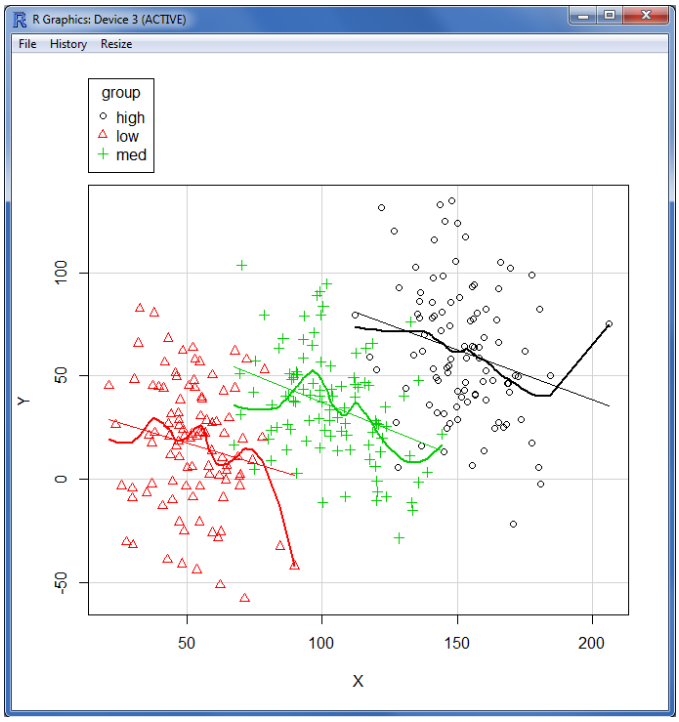
The second example introduces some dependence among the groups and demonstrates the danger of failing to investigate the grouping variable's (Z) influence on the relationship between the primary variables of interest (X & Y).



In the above (script) image, we can see that each of the three groups displays a negative correlation between X and Y (group 1: $r_{xy} = -0.17$; group 2: $r_{xy} = -0.35$; group 3: $r_{xy} = -0.24$). However, if we fail to recognize (i.e. investigate) those groups, we see a positive relationship between X and Y ($r_{xy} = 0.41$) when the groups are not taken into account.

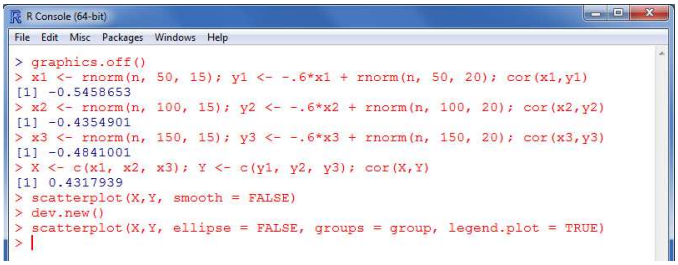


As can be seen above, the positive (overall) relationship between X and Y ($r_{xy} = 0.41$), even graphically displayed, requires a keen eye and at least some suspicion to realize there may be clusters within the data.

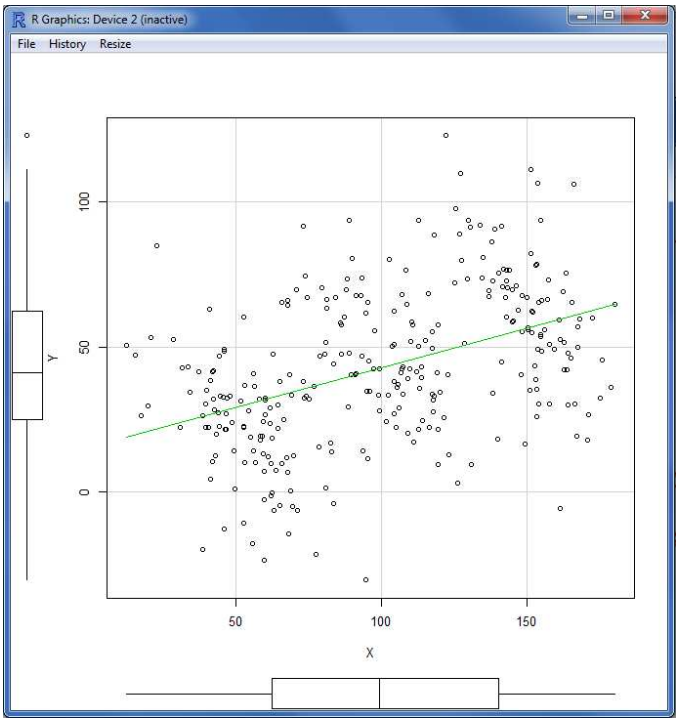


The above image clearly shows the distinctly different nature of the data when the groups are identified. These last two images should clearly demonstrate Simpson's Paradox and the importance of being thorough when conducting initial data analysis. These examples show that it is possible to have the opposite opinion concerning the nature of a relationship between two variables when one does not establish the independence of a third variable. In other words, at first glance X and Y appear to be positively related ($r_{xy} = 0.41$); but once we identify the groups we see that in fact the relationship is negative (group 1: $r_{xy} = -0.17$; group 2: $r_{xy} = -0.35$; group 3: $r_{xy} = -0.24$).

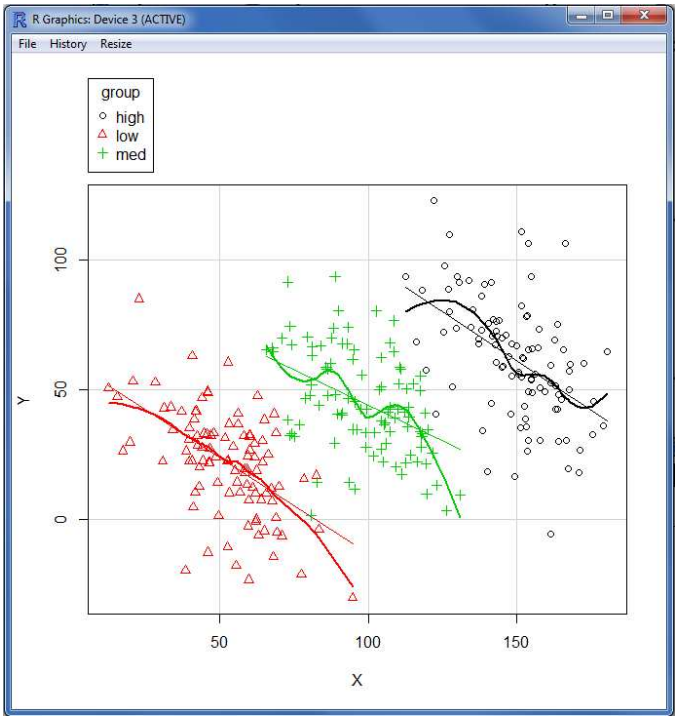
The third example shows a slightly more extreme situation. Below, you can see that the relationship between X and Y is approximately the same as in the previous example ($r_{xy} = 0.43$). However, each group displays a stronger relationship than in the previous example (group 1: $r_{xy} = -0.55$; group 2: $r_{xy} = -0.44$; group 3: $r_{xy} = -0.48$).



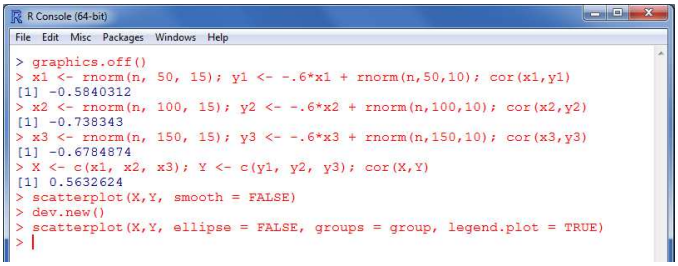
The scatterplot below does not explicitly identify the groups, but given their stronger relationships they are fairly easy to see.



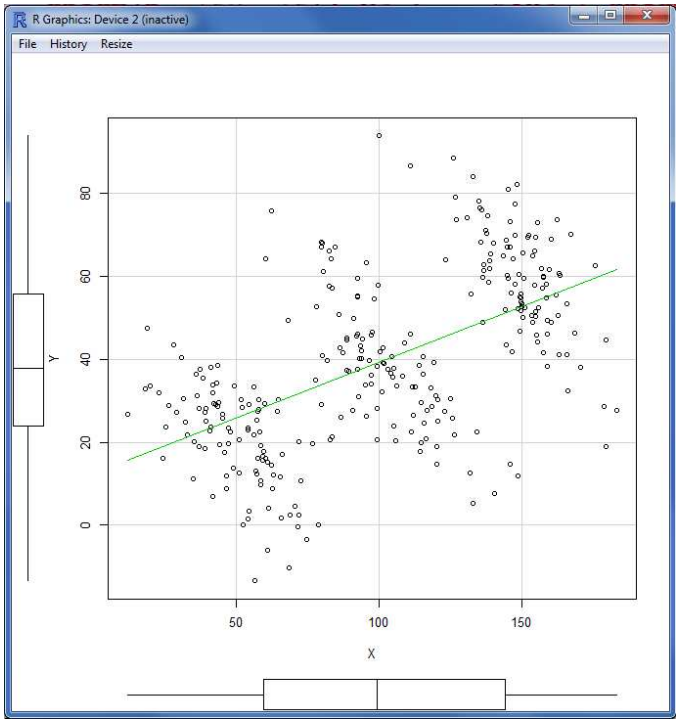
Next, we identify the groups explicitly and plot their (negative) relationships which are clearly in opposition to the overall (positive) relationship.



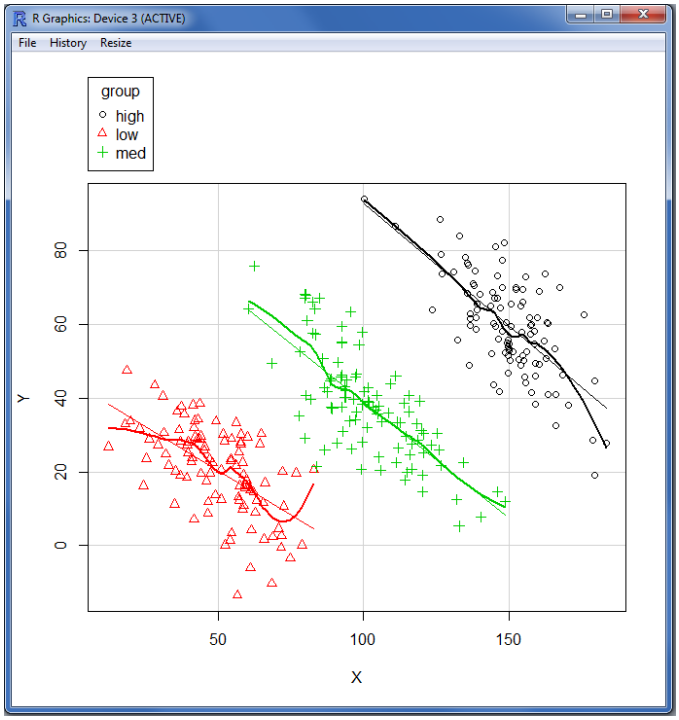
The fourth and final example shows an even more extreme situation. Below, you can see that the relationship between *X* and *Y* is approximately the same as in the previous example ($r_{xy} = 0.56$). However, each group displays a stronger relationship than in the previous example (group 1: $r_{xy} = -0.58$; group 2: $r_{xy} = -0.74$; group 3: $r_{xy} = -0.68$).



This time, the overall scatterplot shows how the clusters of data are distinct enough to be clearly recognizable.



And finally, we identify the groups or clusters explicitly and can see they all display a moderate negative relationship.



The script available on the RSS Introduction to R short course page includes two even more extreme examples. That script is available here.

Conclusions

It is important to recognize that the contrived examples used in this article are extreme in how much different the relationship between X and Y is when accounting for the groups versus not accounting for the groups. Also, in these examples, the group variable was at least recognized and was included in the data collection; often clusters are discovered in patterns of data without prior knowledge of their presence. Therefore, it is extremely important for data analysts to thoroughly investigate and *know* their data intimately. Fortunately, there is a solution (in **R**). Package Simpsons contains a function called 'Simpsons' which tests two continuous variables for the presence of subpopulations (i.e. groups). The function operates by testing whether subpopulations display the same direction and approximate magnitude of relationship as the entire set of cases. The user of the function can supply a suspected grouping variable (e.g., gender / sex) or not. The package also contains functions for summarizing and plotting the results of the 'Simpsons' function. If clusters are recognized in the data, then it may be necessary to collect more / new data simply to explain the clusters. Simpson's paradox is an extreme type of problem, but it should be realized that less extreme situations can (and often do) occur – where the change in relationship may not be as drastic a change as those used in the examples above (i.e. from negative to positive or vice versus). It is also worth noting that even though Robinson (1950) was concerned with clusters among continuous data; Simpson (1951) was interested in contingency tables (i.e. not necessarily continuous variables) and demonstrated the phenomena in that context. Oddly, the paradox is most recognized as Simpson's and not Robinson's.

Until next time; *Happy Festivus...*

References

Bickel, P. J., Hammel, E. A., & O'Connell. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187*, 398 - 404.

Blyth, C. R. (1972). On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association, 67*, 364 - 366.

Clifford, W. H. (1982). Simpson's paradox in real life. *The American Statistician, 36*, 46 - 48.

Fox, J., et al., (2012). Package 'car': A companion package to *An R Companion to Applied Regression*, (2nd Ed.), Sage, 2011.

Kievit, R., & Epskamp. S. (2012). Package 'Simpsons': A package for detecting Simpson's Paradox.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). [Chapter 6: Simpson's paradox, confounding, and collapsibility.] New York: Cambridge University Press.

Ripley, B., Venables, B., Hornik, K., Gebhardt, A., & Firth, D. (2012). Package 'MASS': Functions and datasets to support Venables and Ripley, *Modern Applied Statistics with S* (4th ed.), 2002.

Ripley, B. (2012). Package 'nnet': Software for feed-forward neural networks with a single hidden layer, and for multinomial log-linear models.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351 - 357.

Rucker, G., & Schumacher, M. (2008). Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis. *BMC Medical Research Methodology, 8:34*. DOI: 10.1186/1471-2288-8-34

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (Series B: Methodological), 13*, 238 - 241.

This article was last updated on December 13, 2012.

This document was created using LaTeX