Visualizing multivariate categorical and continuous data from epidemiologic studies: An expanded scatter plot matrix

Benjamin Barnes, Karen Steindorf

German Cancer Research Center, Heidelberg, Germany

Epidemiologic datasets often contain a mix of categorical and continuous variables. Understanding the interrelationships among these variables is vital for subsequent analysis and can be aided by graphical presentation. Scatter plot matrices, produced in R using functions such as pairs() and splom(), are useful for graphically displaying multivariate continuous data. For displaying multivariate categorical data, the Visualizing Categorical Data (vcd) package offers many flexible options, including the pairs.table() function. However, these functions are not readily compatible with one another, making visual presentation of mixed epidemiologic data difficult. With this in mind, the scope of the splom() function was expanded to include visualization of categorical data. Furthermore, a novel panel function compatible with splom() was created to visualize categorical-categorical data using a mosaic plot. Continuous-continuous data was plotted using existing scatter plot and level plot panel functions. Existing panel functions were also used to produce box-and-whisker plots for categorical-continuous data as well as stacked bar charts for categorical-categorical data. With these modifications and the new mosaic function, categorical and continuous data can be viewed in a unified plot matrix. An example of such a plot matrix was created using simulated data inspired by a study investigating the effects of lifestyle and anthropometric factors on insulin-like growth factor (IGF)-I and IGF binding protein (IGFBP)-3. These two proteins are suspected of playing a role in breast cancer development, and current research focuses on identifying modifiable lifestyle factors that influence their concentrations in blood. The expanded scatter plot matrix described here improves visualization of mixed datasets and can be further enhanced to visualize bivariate linear models, chi-squared tests, and other bivariate statistical test results.