

GRAPHICAL DISPLAYS FOR BIVARIATE DATA

William G. Jacoby
ICPSR and Michigan State University

ICPSR Summer Program
July 27-29, 2009

<http://polisci.msu.edu/jacoby/icpsr/graphics>

I. Scatterplot is the Basic Graphical Display for Bivariate Data

- A. Two-dimensional display, showing joint distribution of observations on two variables
- B. Very useful because it avoids the implicit assumptions of statistical models

II. General Guidelines

- A. Use axes on all four sides to enclose the plotting region
- B. Clear axis labels
- C. Rectangular grid lines within the plotting region are usually unnecessary
- D. Plotting symbols should be visually prominent and resistant to overplotting
- E. Tick marks should point outward, rather than inward and relatively few tick marks should be used on each axis
- F. Data rectangle should be slightly smaller than the scale rectangle
- G. Scale rectangle versus data rectangle
- H. If necessary, transform data values (alternatively, used transformed scales on coordinate axes) so plotted points fill up as much of the data rectangle as possible

III. Enhancements for Scatterplots

- A. *Jittering* for alleviating overplotting and dealing with repeated data points
- B. Marginal displays of univariate information
- C. Labeling data points— use labels sparingly!

IV. Slicing a Scatterplot

- A. Scatterplots are usually used to examine functional dependence between variables
- B. Visual assessment of functional dependence in “raw” data is problematic
- C. Dividing the plotting region of the scatterplot into a series of vertical “slices” enables visual display of conditional Y distributions
- D. It is useful to employ a univariate graphical display within each slice
- E. Compare salient characteristics across conditional Y distributions to assess functional dependence

V. Scatterplot Smoothing

- A. Simplest type of nonparametric regression— examine the relationship between two variables without specifying functional form in advance of the analysis
- B. Generate a smooth curve that follows center of conditional Y distribution, across the range of X values
- C. Many methods exist for nonparametric regression and scatterplot smoothing; Loess (or Lowess) is most popular and has nice properties

VI. Loess, or *Local regression*

- A. Loess carries out a series of regressions within a “window” that “moves” across the range of X values.
- B. Define m equally-spaced locations (called v_j , with j ranging from 1 to m) along the range of X
- C. At each v_j , perform a regression, using WLS to estimate the coefficients
- D. In the WLS, observations are weighted inversely with their distance from the current v_j (hence, “local regression”)
- E. The local regression is used to find a predicted value for Y at the current location along the X axis. This is called the fitted (or predicted) value for v_j and is designated $g(v_j)$
- F. Plot the point, $[v_j, g(v_j)]$
- G. Perform m local regressions, calculate fitted values for each, and plot the points, $[v_j, g(v_j)]$ for each of the m locations along the X axis
- H. Adjacent $[v_j, g(v_j)]$ points are connected with line segments to form a relatively smooth curve

VII. Considerations in Using Loess

- A. Smoothing parameter, α , gives proportion of data within each fitting window
- B. Degree of polynomial used in local fitting, λ
- C. Robustness iterations

- D. Loess residuals can be used to guide specification of loess parameters
- E. Residual-fit spread plot for goodness-of-fit (display can also be used for parametric regression models)

VIII. **Aspect Ratio and Banking**

- A. Aspect ratio is the physical size of the vertical scale, relative to the physical size of the horizontal scale
- B. Banking to 45 degrees is the procedure for setting the aspect ratio to maximize differences in slopes of line segments within a smooth curve
- C. Banking optimizes visual perception of functional dependence, and can reveal details otherwise hidden in the graphical display