Learning to Visualize Relationships: A Guide to Creating and Customizing Scatterplots in Stata

Authored by

Mohammed looti

November 8, 2025

The scatterplot stands out as one of the most fundamental and indispensable graphical displays in the field of data analysis. Its core function is to visually map the relationship between two quantitative variables. By positioning paired data points within a Cartesian coordinate system, analysts gain immediate insight into the nature, direction, and strength of the association between the elements under investigation.

Effective visual interpretation of a scatterplot is paramount for successful preliminary statistical exploration. For example, the plot instantly clarifies whether the relationship is **positive** (as the value of one variable rises, the other tends to follow suit) or **negative** (as one variable increases, the other decreases). Furthermore, the density and shape of the data cloud around a potential trend line indicate the strength of the correlation. This initial visual evidence is critical for determining whether complex modeling techniques, such as linear regression, are appropriate and warranted for further study.

This comprehensive guide is designed to serve as an expert resource for creating, customizing, and interpreting scatterplots using the powerful statistical software package, Stata. We will transition from the essential commands needed to generate basic plots to the advanced modifications required to produce graphics suitable for professional publications and rigorous academic reporting.

Preparing Your Data Environment in Stata

Before any visualization can commence, it is essential to properly load and prepare the necessary data within the Stata environment. For instructional purposes throughout this tutorial, we will utilize the widely known, built-in sample dataset named *auto*. This dataset offers rich information on 74 different automobiles, covering key characteristics like weight, length, and displacement, making it an excellent resource for exploring bivariate relationships.

To effortlessly load this standard dataset directly from the Stata Press repository, input the following command into your Stata Command window. This action ensures the data frame is active and ready for subsequent graphical analysis and plotting:

use http://www.stata-press.com/data/r13/auto

Once the dataset is active, it is always best practice to conduct a preliminary inspection. Executing the **summarize** command provides a rapid statistical overview of all variables, including the count of observations, mean, standard deviation, and the minimum and maximum values. This crucial preparatory step confirms successful data loading and helps establish the relevant ranges of the variables we plan to visualize:

summarize

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

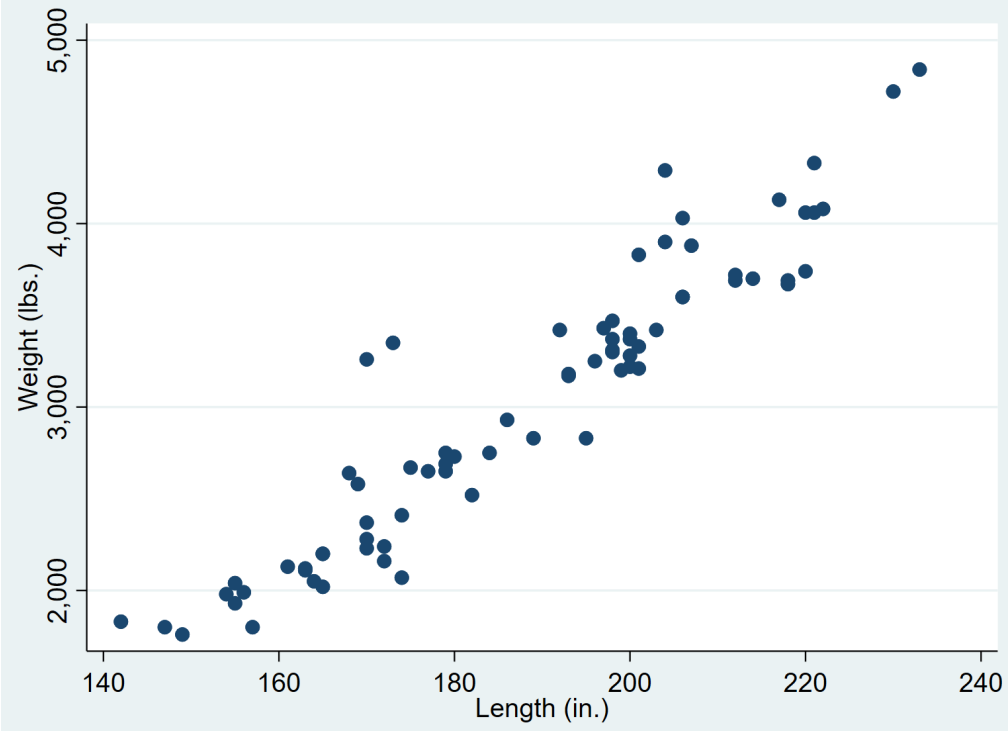
As confirmed by the output above, the *auto* dataset comprises 12 variables, all of which are now readily accessible for graphical exploration. We can proceed directly to examining the initial relationships between specific pairs of these characteristics using the core plotting commands.

Generating Core Scatterplots: Bivariate and Trend Lines

The foundational command for producing a scatterplot in Stata is simply **scatter**. The syntax is specific: the Y-axis variable must be listed first, immediately followed by the X-axis variable. This sequential ordering is vital for ensuring the resulting visual display is correctly oriented and interpreted. We begin by examining the relationship between a car's *weight* (Y-axis) and its *length* (X-axis).

To visualize the fundamental bivariate relationship between these two characteristics, execute the command provided below. The resulting plot will instantly display the distribution of data points across the two dimensions:

scatter weight length

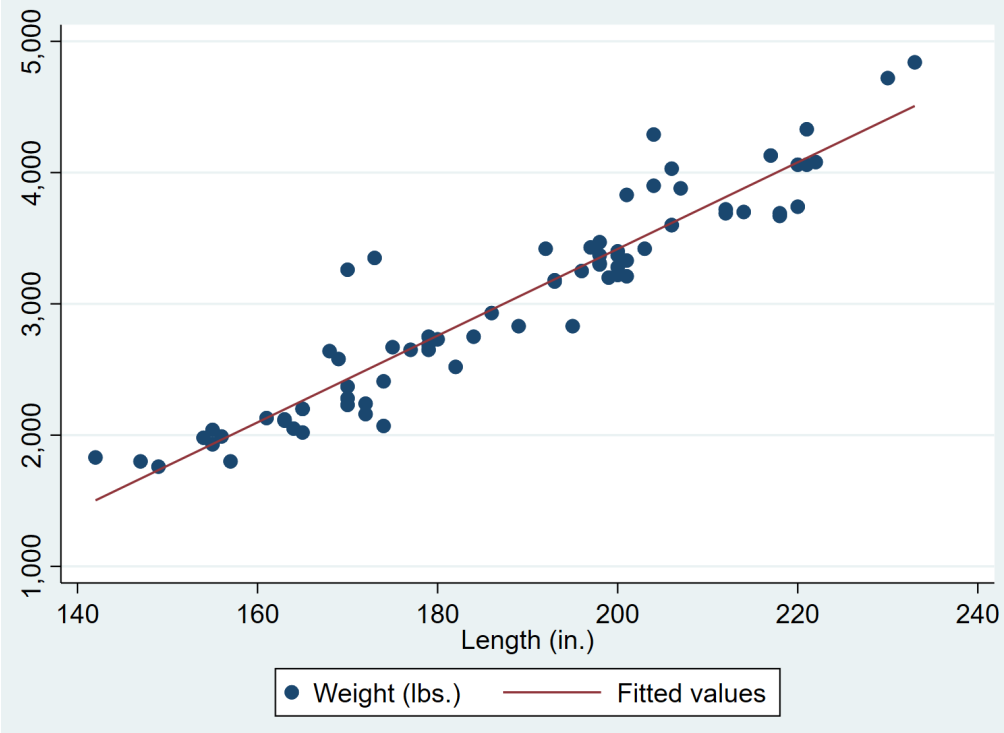


A quick visual inspection of this initial plot suggests a **strong, positive correlation**. This indicates that vehicles with greater weight generally correspond to greater length. Such a clear linear trend frequently serves as the justification for pursuing subsequent quantitative analysis using formal modeling techniques.

Integrating Fitted Regression Lines for Interpretation

For relationships that exhibit a distinct linear pattern, it is tremendously useful to overlay a fitted linear regression line onto the scatterplot. This line, commonly referred to as the "line of best fit," offers a mathematical summary of the average relationship observed within the data. In Stata, this enhancement is achieved by chaining the **scatter** command with the **lfit** command. This combination must be separated by the double pipe symbol (**||**), which is the standard Stata operator used to overlay multiple distinct plots onto a single graph:

scatter weight length || lfit weight length



The resulting graph provides powerful visual confirmation of the linear association, enabling viewers to quickly assess the central average trend and the degree of spread or variance of the individual data points around that summary line.

Visualizing Complexity: Multivariate Plots

Stata offers substantial flexibility by allowing users to plot multiple Y-axis variables against a single X-axis variable on the same graphical display. This capability is exceptionally valuable when the goal is to compare simultaneously how several different metrics relate to one common independent variable. When inputting the variables after the **scatter** command, Stata automatically interprets the very last variable listed as the X-axis variable, while all variables preceding it are treated as separate Y-axis variables.

For example, the command below instructs Stata to create a single scatterplot using *length* as the consistent x-axis variable, and both *weight* and *displacement* as the y-axis variables:

scatter weight displacement length

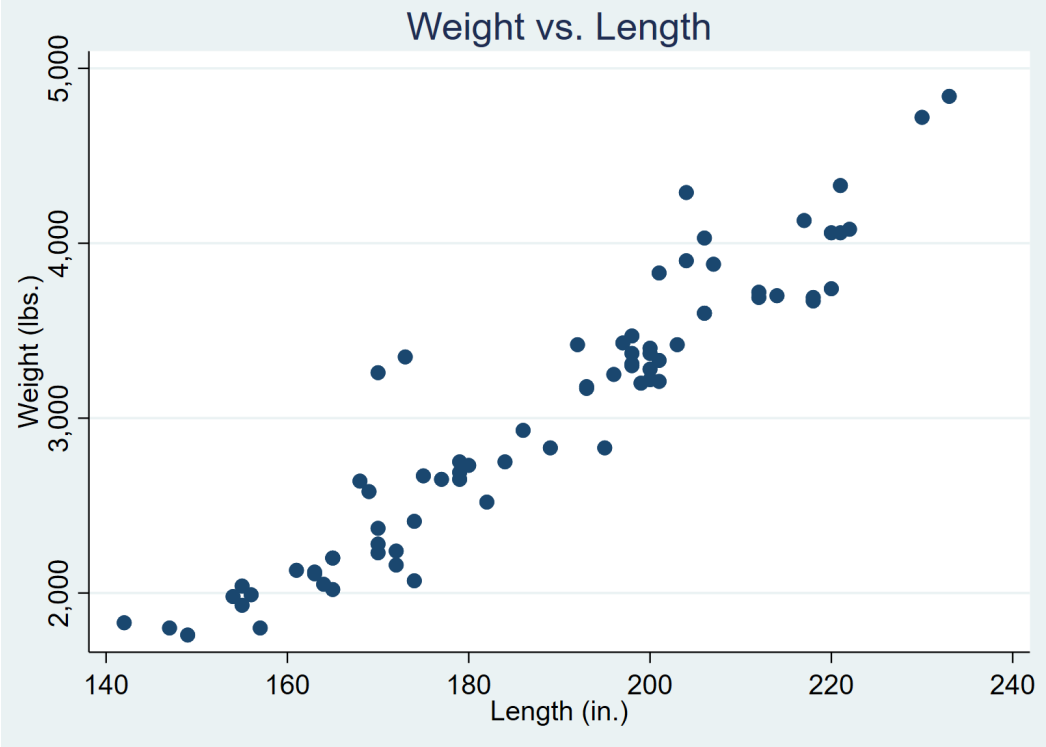
This sophisticated multivariate plot automatically employs distinct symbols and colors for each Y-variable, ensuring that the relationships--in this specific case, between length and weight, and length and displacement--can be observed, differentiated, and compared effectively on the same visualization.

Professionalizing Graphics: Adding Titles and Annotations

While accurately representing the data is the primary objective, effective communication necessitates robust labeling and clear annotation. Modifying the graphical display's textual components--including titles, subtitles, and notes--is crucial for producing publication-ready figures that are fully self-explanatory, regardless of their context. Stata implements these necessary modifications through specific options appended after a comma (,) following the main scatter command.

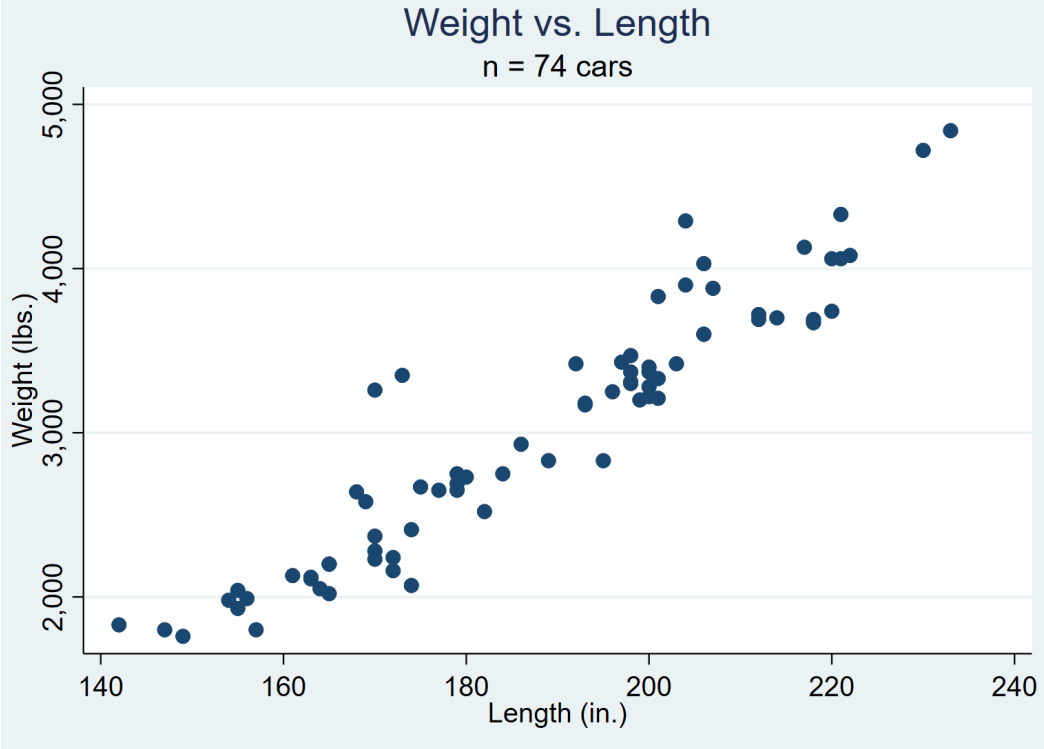
The most immediate way to identify the content of a graph is by including a clear, descriptive title. This is accomplished using the **title()** option, where the desired text is carefully enclosed within quotation marks. A descriptive title instantly informs the audience about the variables and context being displayed:

scatter weight length, title("Weight vs. Length")



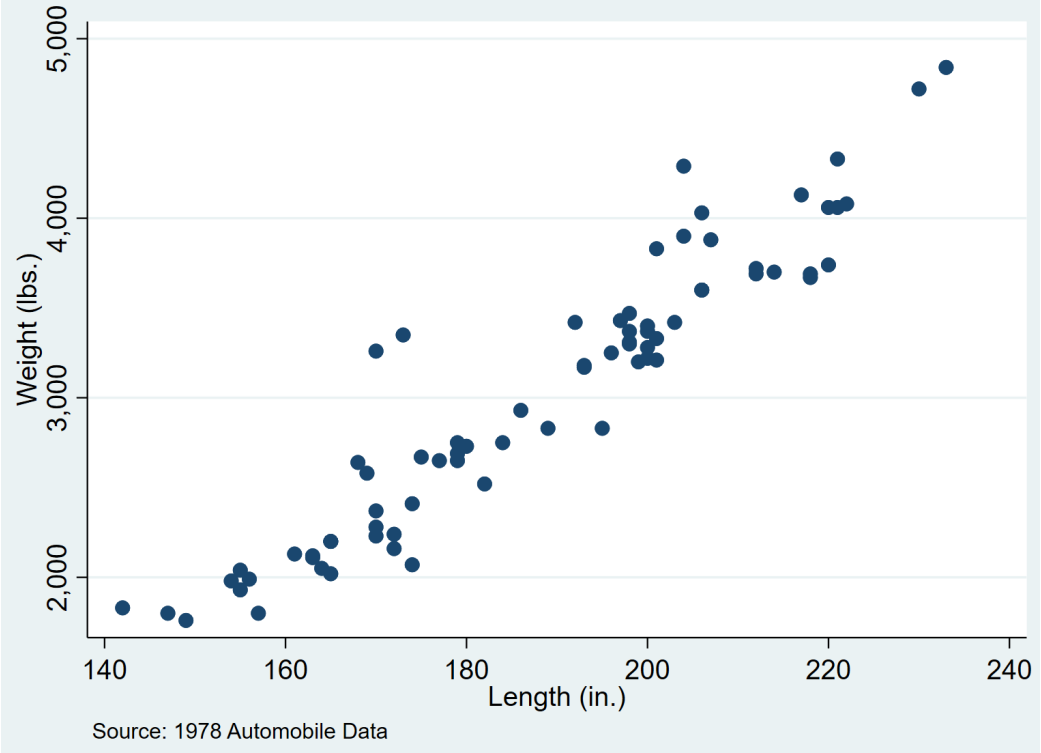
For graphs that require additional context or finer detail that might otherwise clutter the main title, the **subtitle()** option proves invaluable. This feature allows researchers to include supplementary information, such as the sample size (N), details regarding data limitations, or reference to specific sub-groupings, all positioned neatly just below the main title:

scatter weight length, title("Weight vs. Length") subtitle("n = 74 cars")



Finally, professional graphs often require a footnote or a source citation to provide crucial transparency regarding the data's origin or any relevant methodological caveats. The **note()** option positions text at the bottom margin of the graph, functioning as an essential annotation feature. This step is particularly vital when presenting findings derived from external, proprietary, or aggregated datasets:

scatter weight length, note("Source: 1978 Automobile Data")

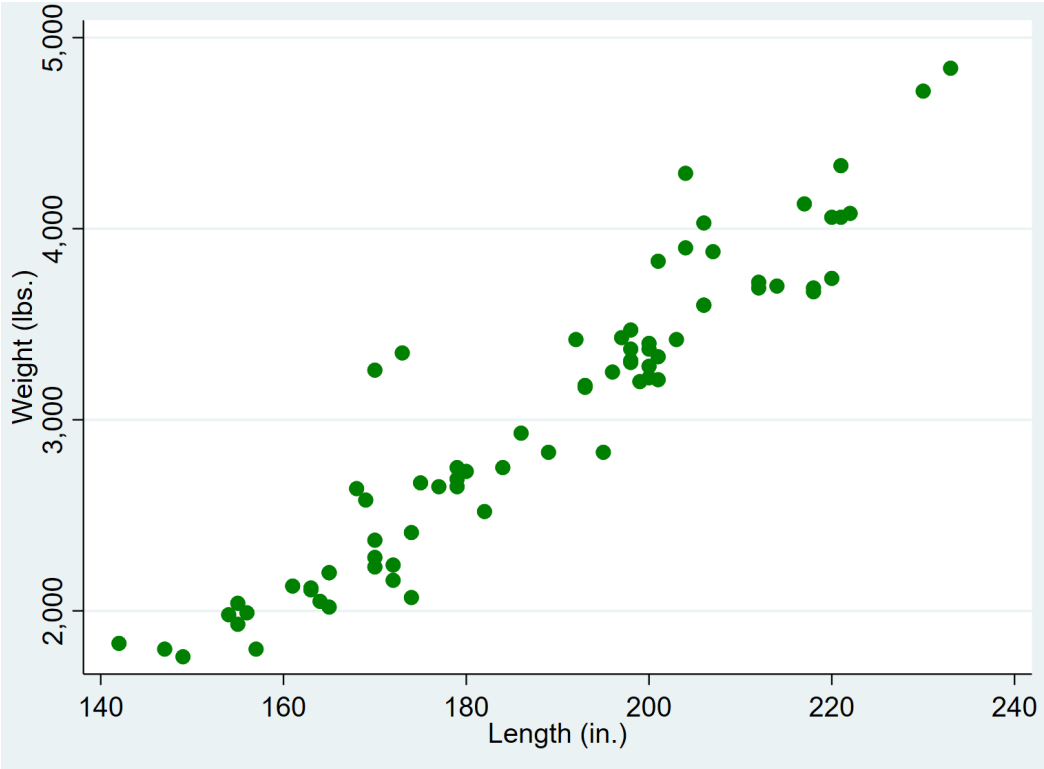


Mastering Aesthetics: Customizing Marker Styles

Beyond textual enhancements, the visual impact and clarity of a scatterplot depend significantly on the aesthetic choices applied to the data markers. Customizing the color and shape of these points is often necessary to adhere to institutional style guides, improve differentiation in complex multivariate plots, or simply enhance the overall readability of the figure.

Although the default marker color in Stata is typically a standard blue, this can be easily modified using the **mcolor()** option, which controls the marker color. Stata supports a wide spectrum of standard color names and specific RGB specifications, granting precise control over the visual palette. For example, to change the markers to a vibrant green, the syntax is straightforward:

scatter weight length, mcolor(green)

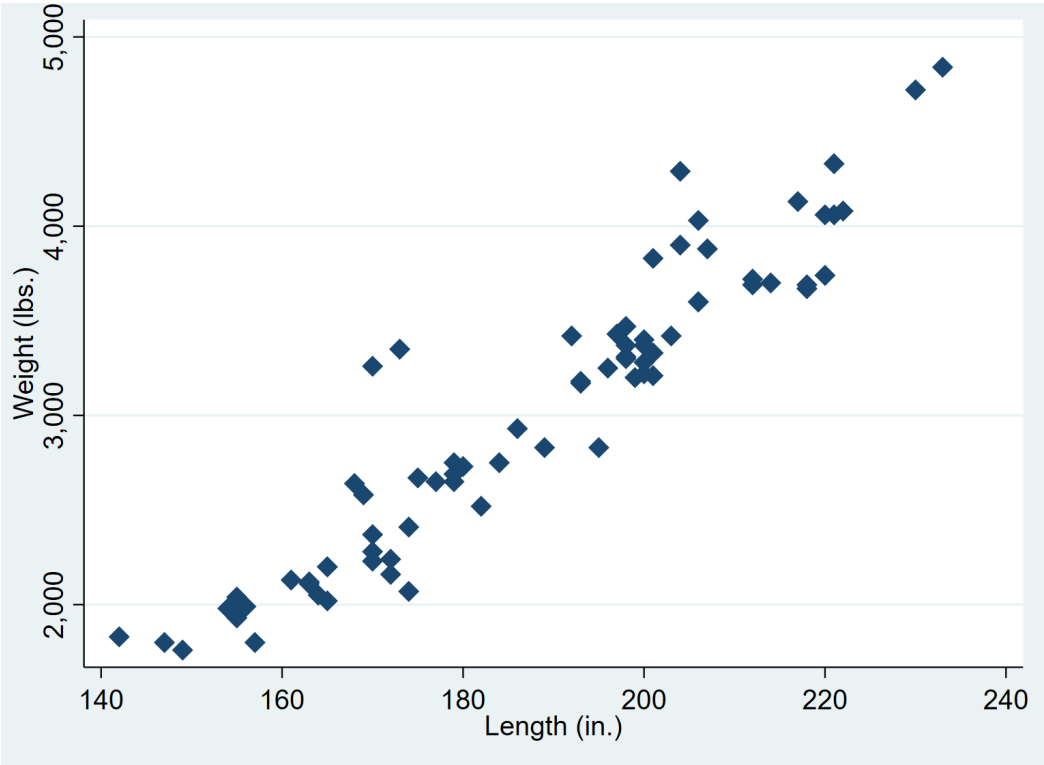


For researchers demanding specific hues or complex palettes, consulting the official Stata Graphics Manual is highly recommended, as it provides a complete inventory of available color schemes and advanced customization methods essential for meeting professional output standards.

Similarly, the default marker shape is generally a solid circle, but this can be altered using the **msymbol()** option, which dictates the marker symbol. Changing the symbol is particularly advantageous when generating overlapping plots or when specific shapes are preferred for publication clarity. Stata utilizes concise single-letter codes to denote different geometric shapes.

For instance, if the analysis requires the individual data points to be represented by diamonds instead of the default circles, the code "D" is utilized within the option:

scatter weight length, msymbol(D)



Like the color options, a comprehensive listing of all available marker symbols (which include triangles, squares, hollow shapes, and more) and their corresponding codes is meticulously detailed in the Stata documentation, ensuring extensive flexibility for visual customization.

Conclusion and Future Visualization Steps

The scatterplot remains an indispensable foundational tool for preliminary data analysis and for communicating bivariate relationships with maximum effectiveness. Utilizing the powerful and remarkably concise syntax of Stata, analysts can quickly transition from raw data to visually compelling graphical summaries, enabling the rapid identification of crucial patterns such as positive or negative correlation, and allowing for the determination of the overall strength of association between two variables.

We have successfully covered the foundational steps required: loading the data, executing the basic **scatter** command, significantly enhancing interpretation by adding a linear regression line (**lfit**), and managing the complexity inherent in multivariate plots. Furthermore, we demonstrated precisely how aesthetic modifications--including adding informative titles and notes, and changing the marker styles--are essential steps that transform a standard plot into a professional-grade visualization.

For professionals and students seeking to deepen their expertise, exploring Stata's extensive graphical options is highly recommended. Future areas of study should include mastering the

customization of axis labels, adding sophisticated legends, using conditional formatting to color points based on a third categorical variable, and exporting graphs in various high-resolution formats necessary for publication. Achieving mastery of these advanced commands ensures that data visualizations are not only statistically accurate but also maximally informative and aesthetically polished.

Additional Resources

A full, detailed list of all available colors and shapes, alongside thorough explanations of graphical schemes and advanced customization options, can be found within the official Stata Graphics Reference Manual. This resource is essential for any serious user aiming for publication-quality output.