Create a Scatterplot with Regression Line in SAS

Authored by

Mohammed looti

November 1, 2025

One of the most efficient and robust methodologies for generating high-quality statistical graphics within the **SAS** environment involves the utilization of the **PROC SGPLOT** procedure. This exceptionally powerful analytical tool provides users with the capacity to rapidly produce complex and precise visualizations, notably including the fundamental combination of a **scatterplot** juxtaposed with an estimated **regression line**.

This comprehensive guide is designed to furnish practical, step-by-step examples that illustrate the implementation of this procedure. We will progress systematically from generating a basic plot to creating highly customized and professionally rendered visualizations, thereby ensuring that your **data visualization** efforts are both analytically accurate and visually compelling for reporting and exploratory analysis.

The Analytical Power of Scatterplots and Linear Regression

The SGPLOT procedure is central to the extensive SAS Statistical Graphics (SG) capabilities, having been specifically engineered for the quick and efficient creation of single-cell graphs. When the objective is to analyze and visually represent the quantitative relationship between two continuous variables, the scatterplot serves as the foundational graphic, immediately revealing the direction, form, and strength of any potential correlation.

By adding a fitted regression line, the scatterplot is immediately transformed into a sophisticated analytical instrument. This line represents the mathematical best fit for the observed data, calculated precisely through the established method of least squares. This overlay provides a critical visual interpretation of the estimated **linear regression** model, allowing analysts to immediately gauge the expected change in the dependent variable given a change in the independent variable.

A significant advantage of leveraging SGPLOT is its streamlined and intuitive syntax. This procedure requires minimal coding effort to achieve professional-grade graphic output, making it the ideal choice for performing rapid exploratory data analysis, generating preliminary reports, and creating graphics suitable for immediate dissemination.

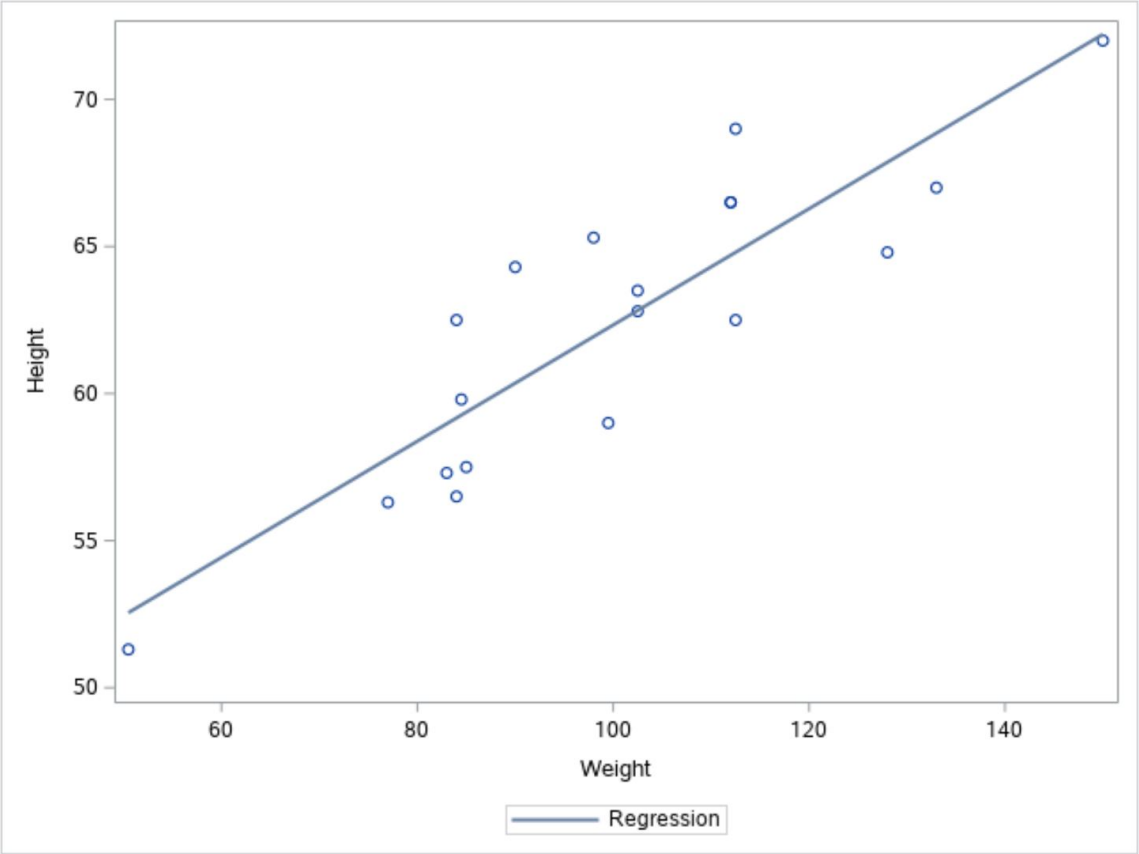
Generating the Foundational Scatterplot in SAS

To commence the visualization process, we will utilize the standard, built-in `sashelp.class` dataset, which is routinely used as a reference dataset within the **SAS** environment and contains essential observations on the height and weight of students. The core command necessary for achieving this visualization is the `REG` statement, which is nested within the **PROC SGPLOT** block. The `REG` statement is highly efficient, as it automatically performs the complex calculations required to derive the regression line and subsequently draws it over the distribution of scattered data points.

The following syntax block illustrates the most simplistic and direct implementation necessary to generate a scatterplot where the variable height is plotted on the Y-axis against weight on the X-axis, complete with the corresponding fitted line:

```
/*create scatterplot with regression line*/
proc sgplot data=sashelp.class;
reg y=height x=weight;
run;
```

Executing this precise code produces the initial graphic, which clearly visualizes the linear relationship between the two variables, establishing the empirical trend based on the input data.



In this visual output, the individual points represent the **observed data points** originating directly from the dataset, illustrating the empirical distribution of height and weight. Crucially, the superimposed blue line is the **fitted regression line**, which is mathematically derived using the method of least squares to minimize the vertical distance between the line and all data points, providing the best linear estimate of the trend.

Achieving Publication Quality: Advanced Customization Techniques

While the basic plot generated above is fully functional for preliminary analysis, **PROC SGPLOT** truly distinguishes itself through its comprehensive capacity to create highly customized visualizations. This customization is essential for generating graphics suitable for professional presentations, academic reports, or publications, ensuring the plot is clear, accessible, and adheres rigorously to specific formatting or branding standards.

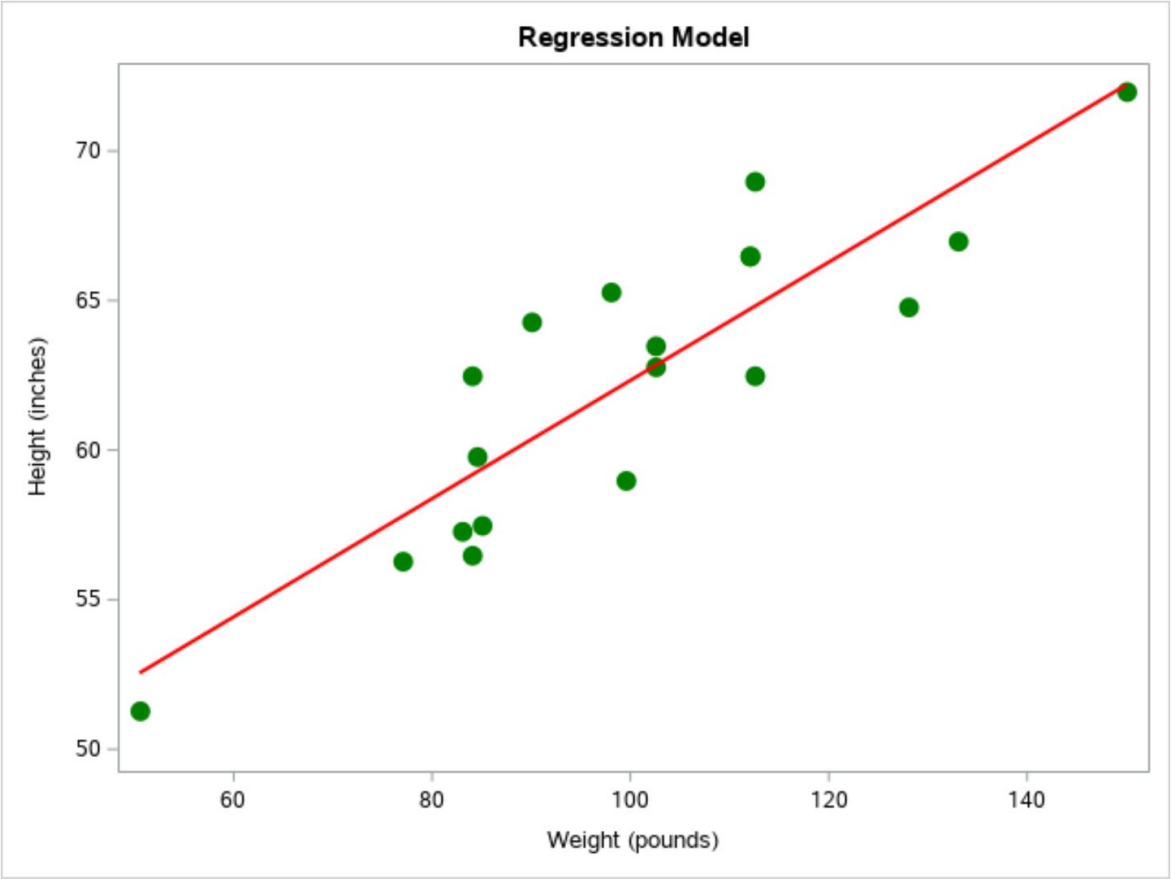
Key customization options are readily available within the SGPLOT procedure when generating scatterplots:

Adding a descriptive title to the chart using the `TITLE` statement, which enhances immediate context and readability for the audience.
Modifying the axis labels using the `XAXIS` and `YAXIS` statements, allowing for precise definition of variable units and contextual information.
Removing the automatic legend using the `NOAUTOLEGEND` option, particularly useful when the plot components are self-explanatory or annotated elsewhere.
Customizing the color, thickness, and style of the regression line itself using the `LINEATTRS` option to ensure aesthetic consistency and emphasis.
Customizing the visual appearance (color, size, symbol) of the individual data points in the plot using the `MARKERATTRS` option for improved differentiation or branding.

The following expanded code block demonstrates how to apply these critical modifications simultaneously. This comprehensive approach results in a polished, fully annotated graphic that is ready for formal reporting:

```
/*create custom scatterplot with regression line*/
proc sgplot data=sashelp.class noautolegend;
title 'Regression Model of Height vs. Weight';
xaxis label='Weight (pounds)';
yaxis label='Height (inches)';
reg y=height x=weight /
lineattrs=(color=red thickness=2)
markerattrs=(color=green size=12px symbol=circlefilled);
run;
```

The resulting visual confirms that the chart title, axis labels, individual data points, and the regression line have all been successfully modified according to the specific attributes detailed in the code. This level of granular control over aesthetic and functional elements is paramount for producing publication-ready statistical graphics.



Visual Interpretation: Understanding Correlation and Model Fit

The primary analytical advantage gained by incorporating a regression line is the immediate visual representation of the correlation between the variables. The slope of the line is directly proportional to the strength and direction of the relationship: a steep slope suggests a strong correlation, indicating a significant change in Y for a unit change in X, whereas a relatively flat line indicates a weak or negligible linear relationship. In the example charting height versus weight, the clear upward slope suggests a strong positive correlation--as weight increases, height tends reliably to increase.

The vertical distance of each observed data point from the fitted line represents the residuals, which essentially quantify the error of the model's prediction for that specific observation. When data points are tightly clustered around the regression line, this pattern indicates a high goodness-of-fit. A tight cluster implies that the linear model accurately captures the underlying structure of the data, minimizing prediction errors.

Furthermore, visually inspecting the scatterplot is a critical step in advanced data analysis, as it aids in identifying potential violations of the core assumptions of **linear regression**. These violations might include patterns of non-linearity (where the trend appears curved rather than

straight) or heteroscedasticity (where the spread of the residuals is unequal across the range of X values), which would necessitate alternative modeling techniques.

Enhancing Rigor: Adding Confidence Intervals and Group Analysis

To substantially enhance the analytical rigor of the scatterplot visualization, **PROC SGPLOT** offers specialized options for including statistical boundaries. The most frequently employed addition is the calculation and display of the **confidence interval**, which serves to quantify the inherent uncertainty associated with the position of the fitted regression line.

To implement the 95% confidence interval (CI) for the mean predicted value, the user simply appends the `/ CLM` option directly to the `REG` statement. This action instructs **SAS** to draw shaded regions surrounding the primary line, visually illustrating the range within which the true population mean relationship between the variables is statistically likely to reside. This graphical representation is vital for understanding the precision of the model's estimation.

Moreover, analysts often require the ability to examine whether the relationship between X and Y differs significantly across specific categorical variables (e.g., comparing males and females). This comparative analysis is easily executed using the `GROUP=` option. By specifying `GROUP=Gender` within the `REG` statement, SGPLOT automatically generates separate, distinct regression lines and corresponding confidence bands for each category, thereby enabling a powerful and immediate comparative **data visualization** of subgroup trends and differences.

Conclusion: Mastering Insightful Data Visualization with PROC SGPLOT

The **PROC SGPLOT** procedure provides a robust, efficient, and flexible framework for the professional creation of scatterplots combined with regression lines in **SAS**. By diligently mastering the core statements, specifically `REG`, and becoming proficient with the customization options such as `LINEATTRS` and `MARKERATTRS`, users can quickly and effectively transform raw observational data into insightful statistical graphics.

These visualizations are essential for clearly communicating complex statistical relationships and supporting evidence-based decision-making. For those seeking to expand their command of statistical plotting and advanced data analysis within the SAS environment, the following resources and tutorials explain how to perform other common tasks and explore additional graphic types: