Calculate Correlation in SAS (With Examples)

Authored by

Mohammed looti

November 1, 2025

In statistical analysis, the primary method used to quantify the linear relationship between two continuous variables is the correlation coefficient. This standardized metric is essential for data scientists and researchers, as it provides a clear measure of both the **strength** and the **direction** of the linear association present in the data. Understanding this relationship is often the first step in building predictive models or determining causality.

The value of the correlation coefficient is always bounded between -1 and 1. This range offers immediate, intuitive insight into how the variables move relative to one another:

-1 signifies a perfectly negative linear correlation. This means that as one variable increases, the other variable decreases consistently and proportionally.
0 indicates that there is no linear correlation between the two variables. They move independently without a discernible linear pattern.
1 represents a perfectly positive linear correlation, where both variables increase or decrease together in lockstep.

Furthermore, the magnitude, or absolute value, of the coefficient dictates the strength of the association. Coefficients closer to 1 or -1 indicate a substantially stronger relationship, suggesting that changes in one variable are highly predictable based on changes in the other. Conversely, values near zero suggest a weak or negligible linear link.

This comprehensive guide will detail the methodology for calculating these crucial coefficients within the powerful SAS software environment. Specifically, we will focus on leveraging the dedicated proc corr procedure. For our practical examples, we will utilize a widely available built-in SAS dataset named **sashelp.Fish**, which contains 159 morphometric measurements collected from fish sampled in a Finnish lake.

Initial Data Inspection: Exploring the Fish Dataset

Before initiating any complex statistical analysis, such as calculating correlation, it is considered best practice to examine the structure and initial observations of the dataset. This preliminary step helps ensure data quality, variable suitability, and correct understanding of the data types. For our purposes, we will inspect the **sashelp.Fish** dataset.

We can efficiently view the first few records of the data using the **proc print** procedure in SAS. The following code snippet instructs SAS to display only the first 10 observations, providing a quick snapshot of the variables available for analysis:

```
/*view first 10 observations from Fish dataset*/
proc print data=sashelp.Fish (obs=10);

run;
```

Obs	Species	Weight	Length1	Length2	Length3	Height	Width
1	Bream	242	23.2	25.4	30.0	11.5200	4.0200
2	Bream	290	24.0	26.3	31.2	12.4800	4.3056
3	Bream	340	23.9	26.5	31.1	12.3778	4.6961
4	Bream	363	26.3	29.0	33.5	12.7300	4.4555
5	Bream	430	26.5	29.0	34.0	12.4440	5.1340
6	Bream	450	26.8	29.7	34.7	13.6024	4.9274
7	Bream	500	26.8	29.7	34.5	14.1795	5.2785
8	Bream	390	27.6	30.0	35.0	12.6700	4.6900
9	Bream	450	27.6	30.0	35.1	14.0049	4.8438
10	Bream	500	28.5	30.7	36.2	14.2266	4.9594

This initial inspection confirms that the dataset contains several critical numeric variables, including **Weight**, **Length** (in multiple forms), **Height**, and **Width**. These variables represent continuous measurements and are thus highly suitable for the calculation of the correlation coefficient, which relies on continuous data for meaningful results.

Example 1: Pairwise Correlation Using the VAR Statement

In many research scenarios, the interest lies specifically in quantifying the relationship between a predefined pair of variables rather than the entire dataset. For this example, we will calculate the Pearson correlation coefficient--the most common measure of linear correlation--between the fish measurements **Height** and **Width**.

To achieve this specific calculation using proc corr, we employ the powerful **VAR** statement. The **VAR** statement explicitly tells proc corr which variables should be included in the analysis, efficiently narrowing the scope to just the two variables of interest:

```
/*calculate correlation coefficient between Height and Width*/
proc corr data=sashelp.fish;
var Height Width;

run;
```

The CORR Procedure

2 Variables:	Height Width

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Height	159	8.97099	4.28621	1426	1.72840	18.95700
Width	159	4.41749	1.68580	702.38020	1.04760	8.14200

Pearson Correlation Coefficients, N = 159 Prob > |r| under H0: Rho=0		
	Height	Width
Height	1.00000	0.79288 <.0001
Width	0.79288 <.0001	1.00000

The output generated by proc corr is typically divided into two distinct sections. The first section provides descriptive summary statistics (N, Mean, Standard Deviation, Minimum, and Maximum) for the included variables (Height and Width), offering context for the data distribution. The second and most critical table, labeled "Pearson Correlation Coefficients," presents the calculated correlation value along with its statistical significance.

Reviewing the results for Height and Width, we observe the following key metrics:

The calculated Pearson correlation coefficient is **0.79288**.
The corresponding P-value is reported as **<.0001**.

This analysis reveals a **strong positive linear correlation** ($R \approx 0.79$) between the height and width measurements of the fish. Crucially, because the P-value is extremely small (far below the conventional significance threshold of $\alpha = .05$), we can confidently conclude that this correlation is statistically significant and not likely due to random chance.

Example 2: Generating a Comprehensive Correlation Matrix

When working with datasets that feature numerous numeric variables, manually running pairwise correlations becomes inefficient. A more streamlined approach is to generate a comprehensive correlation matrix, which calculates the Pearson correlation coefficient for every possible unique combination of variables simultaneously. This matrix serves as an invaluable diagnostic tool, offering a holistic view of multivariate relationships.

In proc corr, calculating the correlation among all numeric variables is remarkably simple. We achieve this by initiating the procedure but intentionally omitting the **VAR** statement. When the **VAR** statement is absent, SAS automatically identifies and processes all appropriate numeric variables within the specified dataset:

/*calculate correlation coefficient between all pairwise combinations of variables*/
proc corr data=sashelp.fish;

run;

The CORR Procedure

6 Variables:	Weight Length1 Length2 Length3 Height Width

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Weight	158	398.69557	359.08620	62994	0	1650
Length1	159	26.24717	9.99644	4173	7.50000	59.00000
Length2	159	28.41572	10.71633	4518	8.40000	63.40000
Length3	159	31.22704	11.61025	4965	8.80000	68.00000
Height	159	8.97099	4.28621	1426	1.72840	18.95700
Width	159	4.41749	1.68580	702.38020	1.04760	8.14200

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Weight	Length1	Length2	Length3	Height	Width
Weight	1.00000 158	0.91644 <.0001 158	0.91937 <.0001 158	0.92447 <.0001 158	0.72869 <.0001 158	0.88741 <.0001 158
Length1	0.91644 <.0001 158	1.00000 159	0.99952 <.0001 159	0.99203 <.0001 159	0.62538 <.0001 159	0.86705 <.0001 159
Length2	0.91937 <.0001 158	0.99952 <.0001 159	1.00000 159	0.99410 <.0001 159	0.64044 <.0001 159	0.87355 <.0001 159
Length3	0.92447 <.0001 158	0.99203 <.0001 159	0.99410 <.0001 159	1.00000 159	0.70341 <.0001 159	0.87852 <.0001 159
Height	0.72869 <.0001 158	0.62538 <.0001 159	0.64044 <.0001 159	0.70341 <.0001 159	1.00000 159	0.79288 <.0001 159
Width	0.88741 <.0001 158	0.86705 <.0001 159	0.87355 <.0001 159	0.87852 <.0001 159	0.79288 <.0001 159	1.00000 159

The resulting output is a densely informative correlation matrix. Each cell in this matrix provides three key pieces of information: the correlation coefficient itself, the sample size (N) used for that specific pair, and the corresponding P-value, which aids in assessing statistical significance for each relationship.

A quick examination of this matrix allows us to efficiently analyze relationships involving the fish's **Weight** and its various length measurements (Length1, Length2, and Length3). The findings confirm a remarkably consistent and strong association:

The Pearson correlation coefficient between **Weight** and **Length1** is **0.91644**.
The Pearson correlation coefficient between **Weight** and **Length2** is **0.91937**.
The Pearson correlation coefficient between **Weight** and **Length3** is **0.92447**.

These exceptionally high positive coefficients--all hovering around R = 0.92--clearly demonstrate that the fish's weight is strongly and linearly related to its length, irrespective of which specific length measurement is used. This kind of redundancy often suggests multicollinearity if these variables were to be used in a regression model.

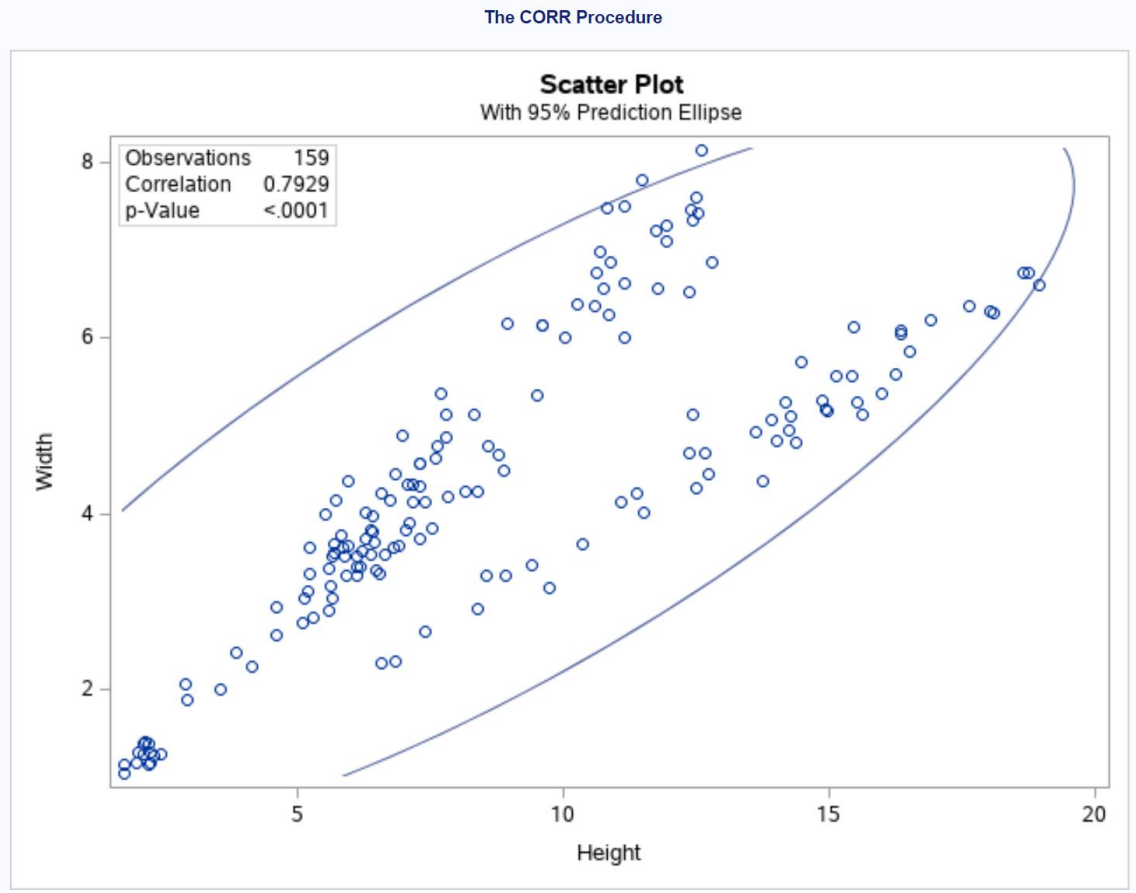
Example 3: Graphical Confirmation Using Scatterplots

While numerical coefficients provide precise measures of association, relying solely on numbers can be misleading, especially if the relationship is non-linear or masked by outliers. Therefore, visualizing the relationship using a scatterplot is a critical step to confirm the assumption of linearity and to visually assess the strength and direction of the correlation.

The **proc corr** procedure in SAS is highly flexible and allows users to generate high-quality graphical output directly through the **PLOTS** option. To confirm the strong positive correlation observed between Height and Width, we will re-run our pairwise analysis and append the necessary plotting statement: `plots=scatter(nvar=all)`. The `nvar=all` sub-option ensures that a scatterplot is created for every pair of variables specified in the **VAR** statement.

```
/*visualize correlation between Height and Width*/
proc corr data=sashelp.fish plots=scatter(nvar=all);;
var Height Width;

run;
```



The resulting scatterplot provides clear visual validation of our numerical findings. The data points form a distinct, tight cloud that slopes upward from left to right, visually confirming the strong positive linear correlation (R = 0.79288). The lack of severe curvature or isolated outliers reinforces the reliability of the calculated correlation coefficient.

A useful feature of this graphical output is the inclusion of key statistical summaries directly on the plot. In the top-left corner, analysts can quickly reference the total number of observations, the exact value of the correlation coefficient, and the associated P-value, integrating both the visual and numerical evidence of the relationship.

Summary and Additional Resources for SAS Procedures

The correlation matrix and individual correlation coefficients calculated using **proc corr** are indispensable tools for initial data exploration and statistical modeling in SAS. By effectively utilizing the **VAR** statement for specific pairs or omitting it to generate a full matrix, researchers can quickly and accurately assess the linear associations between their variables, confirming data patterns visually through the **PLOTS** option.

Understanding how to calculate the correlation coefficient is just one aspect of working with large

datasets in the SAS environment. To continue developing your statistical programming skills, explore other powerful procedures essential for data preparation, manipulation, and advanced modeling: