How to Create and Interpret Q-Q Plots in Stata for Data Analysis

Authored by

stats writer

December 28, 2025

A **Q-Q plot**, short for **Quantile**-Quantile plot, is a fundamental tool in statistical analysis, particularly useful for comparing two probability distributions. It achieves this comparison by plotting the **quantiles** of the two datasets against one another. When applied to **regression analysis**, the Q-Q plot serves a critical diagnostic function: determining whether model **residuals** follow a theoretical distribution, most commonly the **Normal Distribution**.

Understanding the visual output of a Q-Q plot is essential for validating the assumptions of many parametric statistical tests. If the data points align closely with the diagonal reference line, it strongly suggests that the distributions are identical. Deviations from this line, especially at the extremes, can flag specific issues such as skewness, heavy tails, or the presence of significant **outliers**.

The Role of Q-Q Plots in Statistical Diagnostics

In the context of standard statistical modeling, especially when performing a **regression analysis**, one of the primary assumptions is that the errors (or **residuals**) are independently and identically distributed following a **Normal Distribution**. The Q-Q plot provides an efficient, visual method to test this crucial assumption, ensuring the validity of hypothesis tests and confidence intervals derived from the model.

This comprehensive tutorial will guide you through the process of generating and interpreting a Q-Q plot using the statistical software package, **Stata**. We will use a practical example to demonstrate how to fit a regression model, calculate the necessary residuals, and assess their distributional properties visually.

Practical Example: Assessing Residual Normality in Stata

To illustrate the generation and interpretation of the **Q-Q plot**, we will utilize the widely available, built-in auto dataset within Stata. Our objective is to first establish a **multiple linear regression** model and subsequently analyze whether the calculated errors meet the assumption of normality, which is crucial for statistical inference.

In this demonstration, we define the vehicle price as our response variable (dependent variable). We will model this price using two explanatory variables (independent variables): mpg (miles per gallon) and displacement. The subsequent steps focus entirely on extracting the **residuals** from this fitted model and subjecting them to the visual normality test provided by the Q-Q plot.

Step 1: Loading and Summarizing the Data

Before proceeding with model fitting, it is standard practice to load the necessary dataset into the Stata environment and perform an initial inspection. The auto dataset is readily available and

contains various attributes of 74 automobiles. We load this dataset using the sysuse command:

sysuse auto

Following the successful load, a quick summary of the variables is beneficial to understand the distribution, means, and standard deviations of the variables we intend to use in the regression model. This is achieved using the standard summarize command:

summarize

```
. sysuse 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

Step 2: Fitting the Multiple Linear Regression Model

With the data loaded and reviewed, the next logical step involves fitting the specified **regression analysis** model. As previously established, we are modeling price as a function of mpg and displacement. The core command for running an ordinary least squares (OLS) regression in Stata is regress.

Executing this command provides a comprehensive output detailing the model coefficients, standard errors, R-squared values, and F-statistics. This output confirms the relationship between our chosen predictors and the response variable, setting the stage for the diagnostic phase focused on residuals.

regress price mpg displacement

```
. regress price mpg displacement
```

Source	SS	df	MS	Number of obs	=	74
				F(2, 71)	=	13.35
Model	173587098	2	86793549.2	Prob > F	=	0.0000
Residual	461478298	71	6499694.33	R-squared	=	0.2733
				Adj R-squared	=	0.2529
Total	635065396	73	8699525.97	Root MSE	=	2549.4

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-121.1833	72.78844	-1.66	0.100	-266.3193	23.95276
displacement	10.50885	4.58548	2.29	0.025	1.365658	19.65203
_cons	6672.766	2299.72	2.90	0.005	2087.254	11258.28

Step 3: Calculating and Storing the Model Residuals

The **residual** for any observation is defined as the vertical distance between the actual observed value of the response variable and the value predicted by the fitted regression line. Essentially, the residual represents the unexplained variation in the model. Accurate residual calculation is mandatory before we can assess the normality assumption.

In Stata, the predict command is used immediately after running the regress command to generate new variables based on the results of the model. By using the residuals option, we instruct Stata to calculate these error terms for every observation. We store these values in a new variable named resid_price for easy subsequent use:

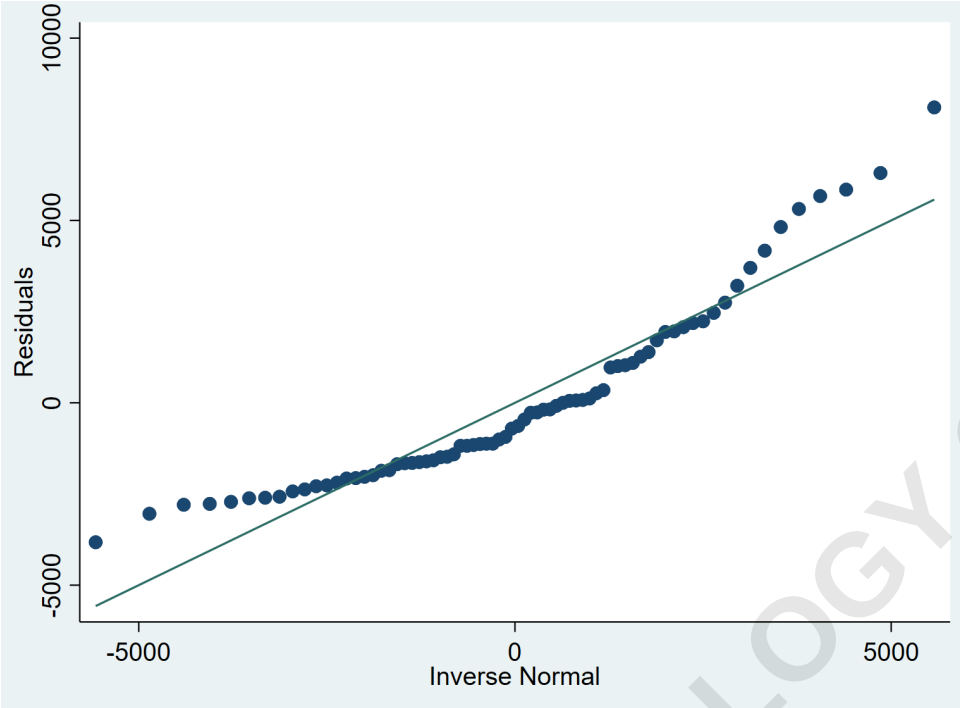
predict resid_price, residuals

Step 4: Generating the Q-Q Plot in Stata

Once the residuals have been computed and stored, we are ready to visualize their distribution using the **Q-Q plot**. Stata simplifies this process significantly through the dedicated qnorm command. This command generates a normal quantile plot, comparing the empirical **quantiles** of the specified variable (in our case, resid_price) against the theoretical quantiles expected from a perfect **Normal Distribution**.

By executing the qnorm command followed by the name of our residual variable, Stata produces the graphical representation necessary for our diagnostic check. The resulting graph plots the ordered residual values against the corresponding standard normal **quantiles**, providing a direct visual test of normality:

qnorm resid_price



Step 5: Interpreting the Q-Q Plot for Normality

The primary principle underlying the interpretation of a normal Q-Q plot is straightforward: if the observed data points are drawn from a **Normal Distribution**, the plotted points should fall approximately along a straight, 45-degree reference line. This line represents where the data **quantiles** would theoretically sit if the distribution were perfectly normal.

Examining the plot generated in Step 4, we observe clear deviations from the idealized straight line, particularly at the extreme ends (the tails) of the distribution. In the lower tail, the points fall below the line, and in the upper tail, they rise above it. Such pronounced bending suggests that our residuals are not strictly normally distributed; specifically, the distribution appears to have heavier tails than a standard normal curve.

Addressing Deviations from Normality

While the **Q-Q plot** is an indispensable visual diagnostic tool, it is important to remember that it does not constitute a formal statistical hypothesis test. It provides qualitative evidence regarding the distributional shape. If the deviations are severe, indicating a major failure of the normality assumption, subsequent steps are necessary to ensure the reliability of the **regression analysis** results.

For large deviations, statisticians often recommend applying a mathematical transformation to the response variable. Common transformations include the square root, the logarithm (log), or the reciprocal. These techniques can often normalize the distribution of the errors, allowing the model to satisfy the underlying assumptions more closely. The specific choice of transformation often depends on the type of skewness observed in the plot.

However, it is crucial to note that **regression analysis** is generally considered robust to minor departures from normality, especially when dealing with large sample sizes (due to the Central Limit Theorem). If the plotted points only show slight wiggles around the 45-degree line, intervention is usually unnecessary, and the model results can be trusted without requiring complex data transformation.