Learning Guide: Understanding and Generating Q-Q Plots in Stata

Authored by

Mohammed looti

November 8, 2025

The **Quantile-Quantile plot** (or Q-Q plot) is a fundamental graphical technique in statistical diagnostics, serving as an indispensable tool for comparing the probability distribution of a specific dataset against a theoretical distribution. In the vast majority of cases, particularly within the framework of linear modeling, this comparison is made against the **normal distribution**. Within the context of **regression analysis**, the Q-Q plot is specifically deployed to visually ascertain whether the model's **residuals** are approximately normally distributed--a crucial assumption required for accurate statistical inference.

The assumption that model errors are normally distributed constitutes a core tenet of classical linear regression. Significant deviations from this requirement can potentially invalidate standard statistical tests, such as t-tests and F-tests, and compromise the integrity of the confidence intervals derived from the estimated model. Consequently, mastering the proper generation and interpretation of Q-Q plots, especially when using powerful statistical software like **Stata**, is paramount for ensuring robust and trustworthy data modeling outcomes. This comprehensive tutorial provides a detailed, step-by-step guide on how to generate and correctly interpret this essential diagnostic visualization using standard Stata commands.

The Essential Role of Q-Q Plots in Statistical Diagnostics

Before proceeding to the practical implementation steps, it is vital to establish a clear conceptual understanding of why we must rigorously examine the distribution of **residuals**. In any standard multiple linear regression model, we proceed under the assumption that the error term--which represents the unexplained variance in the response variable--is independent and identically distributed (I.I.D.) following a normal distribution with a mean of zero and constant variance. If these underlying errors are not **normally distributed**, particularly if they exhibit pronounced skewness or high kurtosis (heavy tails), the calculated standard errors of the coefficient estimates may be inaccurate. This inaccuracy can subsequently lead to distorted t-statistics and potentially incorrect conclusions regarding the statistical significance of the predictor variables.

Violations of the normality assumption, when severe, often signal deeper underlying issues, such as the presence of influential outliers, unaddressed heterogeneity in the sample data, or fundamental model misspecification. While linear regression is often considered robust against minor departures from normality--a property often attributed to the powerful effects of the **Central Limit Theorem**, particularly with large sample sizes--severe non-normality demands immediate attention and may necessitate remedial actions, such as data transformation. The **Q-Q plot** offers the most intuitive and direct visual check for this critical assumption, providing a graphical comparison between the empirical quantiles of our calculated residuals and the theoretical quantiles expected under a perfect standard normal distribution.

Preparing the Data in Stata: The Automobile Example

To effectively demonstrate the entire procedure for generating a Q-Q plot, we will employ a practical example utilizing the built-in *auto* dataset, which is readily available within **Stata**. This dataset contains comprehensive characteristics of numerous automobiles and allows us to construct a realistic multiple regression scenario. Our analytical objective is to fit a multiple linear regression model where *price* acts as the response variable, and *mpg* (miles per gallon) and *displacement* are used as explanatory variables. The initial step is always to load and verify the structure of the data before any analysis can commence.

The process begins by loading the necessary dataset into the current **Stata** session. We achieve this using the `sysuse` command, which is specifically designed to retrieve standard example datasets bundled with the software. Once the data is loaded, standard practice mandates examining the data structure using the `summarize` command. This command provides quick descriptive statistics--such as counts, means, and standard deviations--which ensures that all variables intended for the analysis are correctly loaded and ready for the subsequent modeling stage.

Step 1: Load and view the data summary.

We first load the automobile data using the following command:

sysuse auto

Next, we obtain a quick overview of the dataset variables using the 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

The resulting summary output confirms the availability and structure of all variables required for our **regression analysis**, providing confidence that the necessary fields (price, mpg, displacement) are present and appropriately formatted for immediate use.

Executing the Multiple Linear Regression Model

With the data successfully prepared and loaded into memory, the next essential step is fitting the specified **multiple linear regression** model. Our hypothesized model seeks to utilize *mpg* and *displacement* to predict the dependent variable, *price*. This modeling step is critically important because the accurate subsequent calculation of the **residuals**--defined as the deviation between the observed prices and the prices predicted by the model--is entirely contingent upon the parameter estimates generated during this fitting process.

In **Stata**, the `regress` command is the primary function used for executing **Ordinary Least Squares** (OLS) models. When using this command, the analyst must specify the response variable first, immediately followed by the list of independent explanatory variables. The output generated by `regress` provides comprehensive results, including the estimated coefficients, their standard errors, t-statistics, and overall model fit statistics (such as R-squared and the F-statistic), which would typically be used for formal statistical inference, provided the underlying assumptions are satisfied.

Step 2: Fit the regression model.

We execute the following command to fit the model predicting price:

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

The displayed regression results confirm the successful estimation of the model parameters. While our current focus is not on interpreting the individual coefficients, recognizing that this step establishes the predicted values ($hat{y}$) is fundamental, as these predicted values are absolutely necessary for calculating the model errors.

Deriving and Storing the Empirical Residuals

Once the regression model has been successfully fitted, the immediate next step required for generating the **Q-Q plot** is calculating and saving the **residuals**. It is essential to remember that a **residual** is simply the arithmetic difference between the observed response value (y_i) and the predicted response value ($hat{y}_i$) derived directly from the estimated regression equation. These residuals serve as the empirical estimates of the unobservable true error terms ($epsilon_i$) and quantitatively represent the portion of the response variable's variance that the constructed model fails to explain. Therefore, it is the distribution of these empirical residuals that we must check for adherence to the normality assumption.

Stata offers the highly versatile `predict` command for executing post-estimation diagnostics. To generate the standard residuals, we use the `predict` command, followed by the desired name for the new variable (in this example, `resid_price`), and then specify the `residuals` option. This command ensures that the newly generated variable correctly contains the difference $y_i - hat{y}_i$ for every single observation used in the regression analysis.

Step 3: Calculate and store the residuals.

We obtain the residuals and store them in a new variable named `resid_price` using the following

command:

predict resid_price, residuals

Executing this command successfully creates a new variable within our dataset, populating it with the residual value corresponding to each observation utilized in the regression estimation. This new variable, `resid_price`, now contains the critical data points necessary to visually assess the **normal distribution** assumption via the Q-Q plot.

Visualizing Normality: Generating the Q-Q Plot using `qnorm`

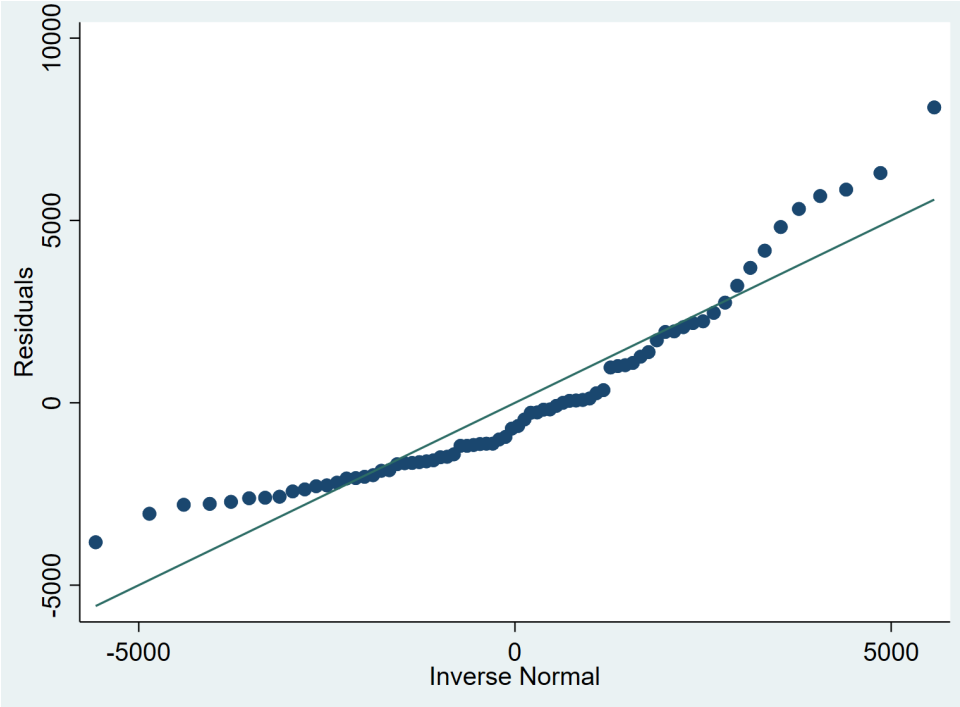
With the **residuals** accurately calculated and stored in the dataset, we can now proceed to generate the actual graphical assessment tool: the **Quantile-Quantile plot**. In **Stata**, the specific command designated for creating a normal quantile plot is `qnorm`. This command automatically executes the necessary calculations to compare the empirical quantiles of the specified variable (our residuals) against the theoretical quantiles that would be expected if the data perfectly followed a standard normal distribution.

The resulting plot graphically displays the observed data points (the residuals) relative to a standardized diagonal line, often referred to as the 45-degree reference line. The x-axis of the plot represents the theoretical quantiles derived from the standard normal distribution, while the y-axis represents the actual observed quantiles of our residual data. If the data points cluster tightly and consistently follow this straight line, it provides compelling visual evidence that the residual data are indeed **normally distributed**. Any noticeable deviations from this reference line, particularly those occurring at the extreme ends or tails, suggest significant departures from the normality assumption.

Step 4: Create the Q-Q Plot.

We utilize the `qnorm` command, specifying our newly created residual variable:

qnorm resid_price



The visual output is instantaneously generated, providing the essential graphical information required for interpreting the underlying distribution of the errors in our **regression analysis**.

Interpreting Results and Addressing Normality Violations

The interpretation of the **Q-Q plot** relies entirely on assessing the proximity of the plotted data points to the straight, diagonal 45-degree reference line. A perfect alignment signifies that the **residuals** adhere perfectly to the theoretical **normal distribution** assumption. Conversely, any observable deviations strongly indicate potential problems with the model's adherence to this critical assumption.

Upon examining the plot generated in Step 4, we can observe that the residuals exhibit a noticeable departure from the 45-degree line, especially evident at the tail ends--both the extremely low and extremely high theoretical quantiles. This specific pattern, where points fall below the line at the lower end and above the line at the upper end, typically suggests that the distribution of our model errors possesses heavier tails than a true normal distribution. This outcome indicates the presence of more extreme residuals (potential outliers) than would be expected under true normality, possibly suggesting a distribution shape closer to a t-distribution rather than a pure Gaussian distribution.

Although the Q-Q plot is not a substitute for a formal statistical test of normality (such as the Shapiro-Wilk test), it offers the quickest and most intuitive method for visually checking the normality of errors. Based on the perceived degree of deviation from the reference line, analysts

must decide on an appropriate course of corrective action:

Addressing Severe Deviation: If the residuals exhibit substantial deviations from the 45-degree line (e.g., pronounced S-shapes, clear curvature, or extreme outlier points), the analyst should seriously consider applying a transformation to the response variable in the regression, such as using the square root, reciprocal, or logarithmic transformation of the response variable (price). Alternatively, researchers may explore more advanced statistical methods, such as robust regression techniques or generalized linear models, which are specifically designed not to rely on the strict assumption of normally distributed errors.

Handling Slight Deviation: If the residuals only diverge slightly from the line, particularly if the deviation is concentrated solely in the center of the plot or if the overall sample size is large, transformation may often be unnecessary. Regression inference tends to be sufficiently robust to minor departures from perfect normality, largely due to the Central Limit Theorem guaranteeing that the distribution of coefficient estimates will approach normality regardless of the error distribution, provided the sample size is adequate.

In summary, the **Q-Q plot** remains an essential and highly effective diagnostic tool for ensuring the statistical reliability and validity of inferences drawn from any linear model. By carefully following these procedural steps within **Stata**, data analysts can rapidly and accurately assess the crucial assumption concerning the distribution of model errors, thereby strengthening the foundation of their statistical conclusions.