How to Run Quantile Regression in Stata: A Step-by-Step Guide

Authored by

stats writer

December 28, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Run Quantile Regression in Stata: A Step-by-Step Guide*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=109492

The **qreg** command is the primary tool for performing Quantile Regression in Stata. This robust command provides researchers and analysts with the ability to move beyond standard estimation of the conditional mean, allowing them to examine various points--or quantiles--of the response variable's distribution. By specifying a quantile of interest, users can execute a regression that focuses precisely on that segment of the data distribution.

Furthermore, the **qreg** command in Stata is highly flexible, supporting advanced analytical features. Users can incorporate additional options to enhance the validity of their results, such as implementing **robust standard errors** to account for heteroskedasticity or utilizing weights to manage complex sampling designs. This flexibility makes it an essential tool for comprehensive statistical modeling, especially when distributional assumptions required by Ordinary Least Squares (OLS) are violated or inappropriate.

1. Understanding the Necessity of Quantile Regression

Linear Regression, often executed using the Ordinary Least Squares (OLS) method, is fundamental in helping us model the relationship between one or more explanatory variables and a single response variable. The core objective of OLS is to estimate the conditional mean of the response variable across the range of the predictors. While incredibly powerful and widely used, OLS focuses exclusively on the average relationship, which may mask crucial insights about the tails or specific segments of the distribution.

In many real-world applications--such as economics, finance, or biological studies--the impact of predictors may vary dramatically depending on where the response variable falls in its own distribution. For instance, factors affecting low-income earners might differ significantly from those affecting high-income earners, even if both groups are part of the same dataset. Relying solely on the mean, as OLS does, provides an incomplete picture. This limitation necessitates a method capable of modeling the entire conditional distribution rather than just its central tendency.

This is precisely where Quantile Regression (QR) distinguishes itself. QR allows researchers to estimate the relationship between predictors and the response variable at various points--or quantiles--of the response distribution. Instead of estimating the mean, we can estimate the median (the 0.50 quantile), the 25th percentile (0.25 quantile), the 90th percentile (0.90 quantile), or any other quantile of interest. This capability provides a much richer, more nuanced understanding of heterogeneous effects across the entire data spectrum.

2. Key Advantages Over Ordinary Least Squares (OLS)

Quantile Regression offers several theoretical and practical advantages when compared to standard OLS estimation. Firstly, QR estimates are robust to outliers in the response variable.

While OLS minimizes the sum of squared errors, which makes it highly sensitive to extreme values, QR minimizes the sum of asymmetrically weighted absolute errors, making it inherently more resistant to non-normal errors and influential observations.

Secondly, unlike OLS which requires the assumption that the conditional variance (heteroskedasticity) is constant across the range of predictors, QR makes no such distributional assumptions regarding the error term. If the error variance is heterogeneous (a common occurrence in economic datasets), OLS estimates can be inefficient, and standard errors unreliable. QR, however, remains consistent and efficient even under these conditions, providing a more reliable analysis when the data exhibit scale effects.

The most compelling advantage is its ability to model **heterogeneous effects**. When performing OLS, we assume the effect of a predictor variable (i.e., the coefficient) is constant across all points of the response variable's distribution. QR relaxes this constraint, allowing the estimated coefficient for a predictor to vary across different quantiles. This is invaluable when the underlying process is complex and non-linear, allowing analysts to uncover relationships that would be entirely invisible using only mean-based estimation.

3. Implementing Quantile Regression Using the qreg Command in Stata

Stata provides straightforward commands for performing Quantile Regression. The main command for single-quantile estimation is **`qreg`**. This command is designed to estimate the parameters of the conditional quantile function, providing coefficients that represent the change in the specified quantile of the response variable for a one-unit change in the predictor.

The basic syntax for the command is simple: `qreg response_variable predictor_variable(s), quantile(q)`. The critical component is the **`quantile()`** option, where `q` represents the quantile level to be estimated, expressed as a fraction between 0 and 1 (e.g., 0.50 for the median, 0.90 for the 90th percentile). If the `quantile()` option is omitted, `qreg` defaults to estimating the median regression (the 0.50 quantile), which is also known as Least Absolute Deviations (LAD) regression.

For more robust inference, it is often advisable to use options available within the **`qreg`** command. For example, researchers often utilize the `vce(robust)` or `vce(bootstrap)` options to obtain heteroskedasticity-consistent standard errors or bootstrapped standard errors, respectively. While QR estimates are robust to outliers in the response variable, proper specification of standard errors is essential for reliable hypothesis testing and confidence interval construction.

4. Example Setup: Loading and Examining the Data

To illustrate the application of Quantile Regression in Stata, we will utilize the popular built-in Stata dataset known as *auto*. This dataset contains comprehensive information on 74 automobiles, covering variables like price, mileage, weight, and engine specifications. Our primary objective in this tutorial is to analyze how the vehicle's **weight** influences its fuel efficiency, measured by **mpg** (miles per gallon), specifically focusing on different parts of the mpg distribution.

We will begin by loading the dataset and performing initial descriptive statistics. This preliminary step is crucial for understanding the range and distribution of our variables of interest: `mpg` (the response variable) and `weight` (the explanatory variable). The following sequence of commands executes these essential preparation steps:

Step 1: Load and view the data.

Use the following command to load the data:

sysuse auto

Use the following command to get a summary of the variables mpg and weight:

summarize mpg weight

```
. sysuse auto
(1978 Automobile Data)

. summarize mpg weight
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mpg	74	21.2973	5.785503	12	41
weight	74	3019.459	777.1936	1760	4840

The summary output confirms the nature of our variables, showing that `mpg` ranges from 12 to 41, and `weight` ranges from 1,760 to 4,840 pounds. This summary provides the foundational statistics necessary before moving into comparative regression modeling.

5. Baseline Comparison: Ordinary Least Squares (OLS) Regression

Before proceeding with Quantile Regression, it is standard practice to establish a baseline model using traditional OLS. This model will estimate the expected conditional **mean** of `mpg` based on `weight`. This allows us to directly contrast the results derived from mean estimation with those derived from quantile estimation later on, highlighting the benefits of the latter when examining non-central tendencies.

Step 2: Perform a simple linear regression.

Use the following command to perform simple linear regression, using `weight` as the explanatory variable and `mpg` as the response variable:

regress mpg weight

```
. regress mpg weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =    134.62
       Model |   1591.9902         1   1591.9902   Prob > F        =    0.0000
    Residual |  851.469256        72  11.8259619   R-squared       =    0.6515
-------------+----------------------------------   Adj R-squared   =    0.6467
       Total |  2443.45946        73  33.4720474   Root MSE        =    3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

The output from the `regress` command allows us to formulate the estimated regression equation for the average car. Specifically, the OLS model suggests a negative relationship: as weight increases, average mpg decreases. The estimated coefficients lead to the following mean prediction model:

We utilize this equation to predict the expected *average* mpg for any given vehicle weight. For example, consider a car that weighs 4,000 pounds. The OLS model estimates its average fuel efficiency to be 15.405 miles per gallon:

predicted average mpg = 39.44028 - 0.0060087*(4000) = **15.405**

6. Performing Single Quantile Regression in Stata

Next, we pivot to Quantile Regression to estimate relationships at a specific point in the distribution. We will focus on the 90th percentile (the 0.90 quantile) of `mpg`. This analysis seeks to answer a different question: What is the estimated fuel efficiency for the **most fuel-efficient 10%** of cars (relative to their weight)? If the relationship between weight and mpg is stronger (i.e., a larger negative coefficient) at the upper quantile, it suggests that weight disproportionately impacts the mileage of highly efficient cars.

Step 3: Perform quantile regression.

We use the **`qreg`** command and specify the desired quantile using the **`quantile(0.90)`** option. This executes the minimization procedure necessary to estimate the parameters for the conditional 90th percentile function:

qreg mpg weight, quantile(0.90)

```
. qreg mpg weight, quantile(0.90)
Iteration  1:  WLS sum of weighted deviations =  80.808183

Iteration  1: sum of abs. weighted deviations =  80.818462
Iteration  2: sum of abs. weighted deviations =  61.133333
Iteration  3: sum of abs. weighted deviations =  55.773684

.9 Quantile regression                          Number of obs =         74
  Raw sum of deviations       90 (about 29)
  Min sum of deviations 55.77368                Pseudo R2     =     0.3803
```

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	-.0072368	.0025281	-2.86	0.006	-.0122766	-.0021971
_cons	47.02632	7.879115	5.97	0.000	31.31959	62.73304

The resulting output provides the coefficients specific to the 0.90 quantile. The estimated regression equation for the 90th percentile is:

predicted 90th percentile of mpg = 47.02632 - 0.0072368*(weight)

Using this QR model, we can calculate the estimated 90th percentile mpg for a car weighing 4,000 pounds. The estimated value is 18.079 mpg:

predicted 90th percentile of mpg = 47.02632 - 0.0072368*(4000) = **18.079**

A direct comparison reveals the utility of QR: the OLS model predicted an average mpg of 15.405, while the QR model for the 90th percentile predicts 18.079. This logical difference demonstrates that a 4,000-pound car requires significantly higher fuel efficiency (**18.079 mpg**) to be considered among the top 10% of efficient vehicles compared to merely achieving the average efficiency (15.405 mpg) for its weight class.

7. Advanced Usage: Simultaneous Quantile Regression (sqreg)

While the **`qreg`** command is excellent for estimating a single quantile, researchers often need to estimate multiple quantiles simultaneously to compare coefficient differences across the distribution. Stata addresses this need with the **`sqreg`** command, which stands for simultaneous quantile regression.

The **sqreg** command allows the user to specify a list of quantiles to be estimated in a single run, often utilizing bootstrapping or robust variance estimators to account for the simultaneous nature of the estimation. This is particularly valuable when the primary research interest is testing whether the coefficients are statistically different across various quantiles, which often signifies important heterogeneous effects.

Suppose we are interested in estimating the 25th percentile (0.25), the median (0.50), and the 90th percentile (0.90) all at once. The syntax requires listing all desired quantiles within the **q()** option:

To do so, we can use the **sqreg** command along with the **q()** option to specify which quantiles to estimate:

sqreg mpg weight, q(0.25, 0.50, 0.90)

```
. sqreg mpg weight, q(0.25, 0.50, 0.90)
(fitting base model)

Bootstrap replications (20)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
....................

Simultaneous quantile regression                    Number of obs =         74
  bootstrap(20) SEs                                 .25 Pseudo R2 =     0.4733
                                                    .50 Pseudo R2 =     0.4934
                                                    .90 Pseudo R2 =     0.3803

------------------------------------------------------------------------------
             |              Bootstrap
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q25          |
      weight |  -.0051724    .0003766   -13.73   0.000    -.0059231   -.0044217
       _cons |   35.22414    1.268633    27.77   0.000     32.69516    37.75311
-------------+----------------------------------------------------------------
q50          |
      weight |  -.0053333    .0005549    -9.61   0.000    -.0064396   -.0042271
       _cons |   36.94667    1.985909    18.60   0.000     32.98783     40.9055
-------------+----------------------------------------------------------------
q90          |
      weight |  -.0072368    .0015364    -4.71   0.000    -.0102997    -.004174
       _cons |   47.02632    5.500986     8.55   0.000      36.0603    57.99233
------------------------------------------------------------------------------
```

This output provides three distinct sets of coefficients. By examining these results, we can construct three separate regression equations, demonstrating how the effect of `weight` changes dramatically across the mpg distribution:

(1) predicted 25th percentile of mpg = 35.22414 - 0.0051724*(weight)
(2) predicted 50th percentile of mpg = 36.94667 - 0.0053333*(weight)
(3) predicted 90th percentile of mpg = 47.02632 - 0.0072368*(weight)

Notice the trend: the coefficient for `weight` becomes increasingly **negative** as we move from the 25th percentile (-0.00517) to the 90th percentile (-0.00724). This statistically confirms that weight has a substantially greater negative impact on the fuel efficiency of cars that are already highly efficient (upper quantiles) compared to those that are less efficient (lower quantiles), reinforcing the necessity of using QR when analyzing distributional effects.