A Practical Guide to Quantile Regression with Stata

Authored by

Mohammed looti

November 8, 2025

Understanding Regression Models: Moving Beyond the Mean

In the realm of statistics and quantitative analysis, the fundamental objective often involves establishing and modeling the relationship between various data components. The most widely employed statistical tool for this purpose is **Linear regression**, a robust technique that allows researchers to quantify the association between one or more **explanatory variables** (or predictors) and a specific **response variable** (or outcome). When analysts utilize standard linear regression, typically implemented through Ordinary Least Squares (OLS) estimation, the primary goal is to estimate the conditional mean of the response variable. This approach provides the expected average value of the outcome, given a fixed set of predictor values. OLS is favored in many disciplines due to its mathematical simplicity, straightforward interpretability, and its optimal statistical properties when core assumptions--such as normality of errors and homoscedasticity--are met.

However, relying solely on the conditional mean introduces significant limitations, particularly when dealing with complex datasets characterized by high volatility, highly skewed distributions, or the presence of influential outliers. Because the mean is inherently sensitive to extreme values, it may fail to accurately represent the typical relationship within a heavily skewed dataset. Crucially, estimating a single average relationship implies an assumption that the influence of the predictors remains constant across the entire distribution of the outcome. This assumption often proves false in real-world scenarios, such as modeling fluctuating market performance, heterogeneous income distribution, or varied environmental impacts, where factors affecting the lowest outcomes differ substantially from those affecting the highest outcomes.

This necessity for a more nuanced modeling approach drives the application of **quantile regression** (QR). Unlike OLS, which focuses exclusively on the center of the distribution (the mean), QR offers the flexibility to estimate the relationship between predictors and various specific locations, known as quantiles, within the conditional distribution of the response variable. This capability enables a researcher to model the effect of covariates at the 10th percentile, the **median** (50th percentile), or the 90th percentile. By examining these diverse points across the distribution, quantile regression provides a significantly more comprehensive and statistically robust picture of how covariates influence the full spectrum of potential outcomes, making it an indispensable technique when distributional heterogeneity is the core subject of inquiry.

Why Choose Quantile Regression over OLS?

The advantages of quantile regression in specific analytical contexts stem from its superior robustness and inherent distributional flexibility. Standard OLS estimation operates by minimizing the sum of squared residuals, a method that mathematically magnifies the impact of large errors associated with outlying data points. In contrast, quantile regression, pioneered by Koenker and

Bassett in 1978, minimizes a sum of absolute residuals that are weighted asymmetrically. This distinct optimization criterion endows QR with robust statistical properties, especially when estimating the conditional **median** (the 0.5 quantile). Consequently, QR estimates are far less susceptible to distortion caused by extreme observations in the response variable compared to estimates derived from traditional OLS procedures.

A second, profoundly critical benefit of QR is its capability to effectively address **heteroscedasticity** without requiring complex remedial measures, such as variance stabilization transformations or elaborate weighting schemes. Heteroscedasticity describes a scenario where the variance of the error term is not constant across all levels of the predictor variables--a common issue in economic and social data. While OLS coefficients remain unbiased under heteroscedasticity, their standard errors become unreliable, potentially leading to erroneous conclusions regarding the statistical significance of the estimated coefficients. Since QR estimates distinct models for each specified percentile, it inherently models the varying scale of the conditional distribution, thereby yielding valid estimates and reliable standard errors even when the variability of the outcome changes drastically in response to the predictors.

Furthermore, QR becomes essential when the primary scientific or business interest is concentrated specifically in the tails of the outcome distribution. For example, in financial risk management, analysts might be focused on accurately modeling the 95th percentile of potential loss (often referred to as Value at Risk). Similarly, in medical research, understanding the factors influencing the slowest recovery rates (represented by the lower percentiles) holds immense practical value. In such specialized cases, OLS provides an estimate averaged over all observations, which is statistically irrelevant to the specific extreme outcome being investigated. Quantile regression empowers the researcher to precisely tailor the model to the exact part of the distribution that carries the greatest analytical weight, providing coefficients that are directly interpretable in the context of that specific quantile boundary.

Practical Implementation in Stata: Baseline OLS

This practical tutorial outlines the execution of quantile regression using the powerful statistical software environment, **Stata**. For demonstration purposes, we will utilize the publicly available, built-in Stata dataset, *auto*, which contains detailed variables related to various car characteristics. Our objective is to model the relationship between a car's `weight` (the explanatory variable) and its fuel efficiency, measured as `mpg` (miles per gallon, the response variable). To establish a foundational comparison, we must first set a baseline using standard linear regression (OLS).

Step 1: Load and View Data. We begin the analysis by loading the required dataset into the **Stata** session and then inspecting the summary statistics for the variables central to our analysis. This crucial initial step confirms that the data has loaded correctly and provides essential context

regarding the central tendency and spread (mean, standard deviation, minimum, maximum) of both `mpg` and `weight` within the sample.

Command to load the data:

sysuse auto

Command to summarize key variables:

summarize mpg weight

The resulting summary output confirms that the analysis utilizes 74 observations and provides the calculated average weight and average miles per gallon for the entire sample population.

```
. 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

Step 2: Perform Simple Linear Regression (OLS). Next, we fit the traditional OLS model using Stata's `regress` command. This benchmark model estimates the expected **average** fuel efficiency (mpg) of a car as a linear function of its weight, serving as the definitive measure for the conditional mean relationship.

Command to perform OLS regression:

regress mpg weight

The output generated by the OLS model provides the estimated intercept and the coefficient corresponding to the weight variable.

```
. 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

Based on this output, we can formulate the estimated conditional mean regression equation:

Average predicted mpg = 39.44028 - 0.0060087 * (weight)

We can then apply this equation to calculate the estimated average mpg for a hypothetical car weighing 4,000 pounds:

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

Analyzing the Quantile Regression Results (90th Percentile)

To truly appreciate the utility of distributional modeling, we now pivot from predicting the average mpg to estimating a high-efficiency performance benchmark. We will execute quantile regression specifically to estimate the 90th percentile of a car's mpg based on its weight (tau = 0.90). This focused analysis targets the relationship governing the most fuel-efficient 10% of vehicles within the sample population.

Step 3: Perform Quantile Regression. In Stata, the appropriate command for single-quantile estimation is `qreg`. We must specify the exact quantile of interest using the `quantile()` option, setting its value to 0.90. This specialized procedure generates coefficients that are uniquely optimized for defining and predicting this high-efficiency boundary, rather than the average.

Command to execute quantile regression for the 90th percentile:

qreg mpg weight, quantile(0.90)

The resulting output displays the estimated coefficients specific to the 0.90 quantile.

```
. 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

Using these newly derived coefficients, we can now establish the estimated regression equation tailored for the 90th percentile:

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

Applying this quantile-specific equation to our 4,000-pound hypothetical car, we determine the estimated mpg required for that vehicle to fall into the 90th percentile of all cars sharing the same weight:

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

The interpretation of this finding is clear and highlights the core benefit of QR: A car weighing 4,000 pounds must achieve a fuel efficiency of 18.079 mpg or greater to be classified among the top 10% most efficient vehicles in that specific weight class. This estimated efficiency benchmark contrasts noticeably with the average predicted mpg of 15.405 derived from the OLS model. Crucially, the QR coefficient for weight (-0.0072368) is steeper (more negative) than the OLS coefficient (-0.0060087). This indicates that the negative impact of increased weight on fuel efficiency is significantly more pronounced when targeting peak performance (the 90th percentile) compared to targeting average performance (the mean). This discovery of a non-constant effect across the conditional distribution is the fundamental insight provided by quantile modeling.

Simultaneous Quantile Regression for Multiple Percentiles

While modeling individual quantiles using `qreg` provides focused insight, a complete and robust distributional analysis frequently necessitates the simultaneous estimation of several quantiles. This approach allows for direct, side-by-side comparison of coefficient estimates across the entire distribution and is essential for conducting formal hypothesis testing regarding the equality of effects--a crucial test for confirming the presence of genuine **heteroscedasticity**. In **Stata**, this multi-quantile analysis is efficiently performed using the `sqreg` command, which stands for Simultaneous Quantile Regression.

The primary methodological advantage of employing `sqreg` is its sophisticated handling of standard error estimation. It estimates the standard errors for the coefficients across all specified quantiles simultaneously, meticulously accounting for the inherent statistical correlation that exists between these quantile estimates. This procedure ensures that the resulting inference is far more accurate and statistically reliable, particularly when performing comparisons, such as testing whether the effect of weight at the 25th percentile differs significantly from its effect at the 75th percentile.

To illustrate, let us estimate the relationships for the lower tail (25th percentile), the absolute center (the **median** or 50th percentile), and the upper tail (90th percentile) in a single operation. We invoke the `sqreg` command and list all three desired quantiles within the `q()` option:

Command to perform simultaneous quantile regression:

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

The resulting output is organized into a single, clean table structure, presenting three distinct regression models that facilitate easy and direct comparison of the intercept and slope coefficients across the chosen quantiles.

```
. 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
```

mpg	Coef.	Bootstrap 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

From this dense and informative output, we can construct the estimated regression equations corresponding to each modeled quantile:

Predicted 25th percentile of mpg = 35.22414 - 0.0051724 * (weight)
Predicted 50th percentile of mpg (Median) = 36.94667 - 0.0053333 * (weight)
Predicted 90th percentile of mpg = 47.02632 - 0.0072368 * (weight)

A careful observation of the weight coefficients reveals a clear and increasing magnitude of the negative effect: -0.00517 (25th percentile), -0.00533 (50th percentile), and a much steeper -0.00723 (90th percentile). This distinct pattern conclusively demonstrates that vehicle weight imposes a substantially greater penalty on the fuel efficiency of high-performing cars than on low-performing cars. Recognizing this non-constant effect across the distribution represents a crucial piece of analytical information that would be completely obscured by relying solely on a single linear regression model.

Conclusion and Further Exploration

Quantile regression constitutes a vital and necessary expansion of the traditional statistical

modeling toolkit. By strategically shifting the analytical focus from merely estimating the conditional mean to precisely modeling specific conditional quantiles, QR empowers analysts to capture the complete, nuanced picture of conditional relationships. This technique provides inherent robustness against influential outliers and yields detailed insights into distributional heterogeneity across the outcome variable's range. It is an indispensable method whenever the impact of an explanatory variable is reasonably suspected to vary depending on the level of the response variable. Fortunately, the implementation using the `qreg` and `sqreg` commands in **Stata** is highly accessible and efficient, transforming sophisticated distributional analysis into a practical routine.

For researchers seeking to deepen their mastery of this method, several advanced considerations are paramount. A key area involves utilizing graphical analysis, specifically plotting the quantile process--the sequence of coefficients estimated across a continuous range of quantiles (e.g., 0.05 to 0.95)--to visually inspect how the effects of predictors change. Furthermore, mastering the correct procedures for statistical inference, which often includes employing bootstrapping methods for reliable standard error estimation, is essential for drawing statistically valid conclusions from quantile regression models. Future exploration might also include applying QR in multivariate settings, incorporating advanced techniques like fixed effects, and utilizing it in complex panel data analysis to further extend its utility.

Additional Resources

To support the advanced utilization and comprehensive understanding of quantile regression in **Stata**, the following resources are highly recommended for acquiring both theoretical depth and practical command syntax mastery:

The official documentation for Stata's `qreg` and `sqreg` commands, which includes detailed explanations of related post-estimation commands necessary for conducting formal hypothesis testing.
Koenker, R. (2005). *Quantile Regression*. Cambridge University Press. This is the seminal work providing the comprehensive mathematical and statistical theory underpinning the method.
User-written Stata commands, such as `grqreg`, which significantly streamline the visualization of quantile regression results and assist in formal testing of coefficient differences across various quantiles.