A Comprehensive Guide to Linear Regression in Stata: Prediction and Residual Analysis

Authored by

Mohammed looti

November 8, 2025

The Foundation of Linear Regression and Diagnostic Tools

Linear regression stands as a cornerstone in statistical modeling, offering a robust framework for understanding and quantifying the relationship between variables. This technique allows analysts to define a linear mathematical relationship between one or more explanatory variables (or predictors) and a single continuous response variable. The fundamental objective is to determine the line of best fit--the equation that minimizes the sum of squared distances between the observed data points and the derived regression line. Achieving this minimum distance ensures the most accurate linear representation of the underlying data trends.

Once a regression model has been successfully estimated using the observed dataset, the resulting equation transforms into a powerful predictive instrument. The calculated coefficients provide the necessary parameters to generate **predicted values** for the response variable, given any specific input combination of the explanatory variables. These fitted values are essential outputs, serving as the primary metrics for assessing the model's predictive power and practical utility in forecasting scenarios. They form the basis for evaluating how well the model generalizes beyond the training data.

Despite the mathematical rigor, no model perfectly captures real-world complexity. Therefore, a critical step in the analysis pipeline involves calculating and analyzing the **residuals**. A residual is formally defined as the vertical distance separating the actual observed value (Y) from the corresponding value predicted by the model ($\hat{Y}$). Specifically, it is calculated as $Y - \hat{Y}$. Analyzing these error terms is central to diagnostic testing, revealing potential systematic errors, identifying outliers, and confirming that the model adheres to the necessary statistical assumptions required for valid inference.

This comprehensive guide outlines the precise steps required within the statistical software package Stata to efficiently generate and systematically analyze both the **predicted values** and the **residuals** following the estimation of any standard linear regression model.

Preparing the Data Environment within Stata

To illustrate the methodology for calculating and inspecting predictive metrics, we will utilize a readily accessible dataset included within the Stata software distribution. This example uses the built-in *auto* dataset, which compiles various characteristics for 74 different automobiles. For our specific regression example, we will model *price* as the primary response variable, attempting to explain its variation using two core predictors: *mpg* (miles per gallon) and *displacement* (engine

size).

The initial operational phase involves loading the sample data into Stata's active memory. This is performed using the dedicated system command designed for accessing built-in examples, followed immediately by a summary command to ensure data integrity and confirm the structure of the loaded variables. This preparatory step is crucial for verifying that the data is correctly structured before any complex statistical estimation begins.

Step 1: Load the dataset and obtain a preliminary summary.

We begin the process by invoking the command to load the designated dataset:

sysuse auto

Following the data load, we execute the `summarize` command to quickly review descriptive statistics--including the count, mean, standard deviation, and range--for all variables present in the dataset:

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

Estimating the Model: Fitting the Regression Equation

With the *auto* dataset successfully loaded and active, the next essential procedure is the formal estimation of the multiple linear regression model. In Stata, this is achieved using the `regress` command, where we must correctly specify the continuous response variable (*price*) followed by the list of selected explanatory variables (*mpg* and *displacement*). This command triggers the calculation of optimal coefficient estimates that define the linear relationship between the predictors and the outcome.

Step 2: Execute the multiple regression command.

The syntax required for fitting our specified model is straightforward:

regress price mpg displacement

The output generated by Stata following the execution of the `regress` command provides a wealth of statistical metrics, including the overall model fit statistics (such as R-squared and the F-statistic) and, most critically, the estimated coefficients for the intercept and each predictor variable. These coefficients are immediately stored in memory, allowing us to formulate the explicit regression equation used for forecasting.

```
. 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

Based on the calculated results, the fitted regression equation can be formally expressed as: $text{Estimated Price} = 6672.766 - 121.1833 times (text{mpg}) + 10.50885 times

(text{displacement})$. It is this derived equation that Stata utilizes internally during subsequent steps to generate all necessary predictive metrics for every observation in the dataset.

Generating and Interpreting Predicted Values

Once the regression coefficients are successfully estimated and reside in Stata's memory, we can leverage the powerful `predict` command to calculate the **predicted values**, also known as the fitted values. These values represent the model's best estimate of the response variable (price) for each observation, based on that observation's specific values for MPG and displacement, according to the established linear relationship.

Step 3: Obtain the predicted values (fitted values).

We instruct Stata to calculate these estimates and store them in a new variable, which we name **pred_price**. When the `predict` command is used immediately after a regression and without any additional options, it defaults to calculating the fitted values ($hat{Y}$):

predict pred_price

To verify the calculation and compare the model's output directly against the raw observed data, we employ the `list` command. Given that the *auto* dataset contains 74 entries, we restrict the output to the first 10 observations using the `in 1/10` qualifier. This side-by-side comparison displays the actual *price* alongside the newly created *pred_price*:

list price pred_price in 1/10

```
. predict pred_price
(option xb assumed; fitted values)

. list price pred_price in 1/10
```

	price	pred_p~e
1.	4,099	5278.305
2.	4,749	7323.933
3.	3,799	5278.305
4.	4,816	6308.834
5.	7,827	8533.113
6.	5,788	6919.01
7.	4,453	6716.69
8.	5,189	6308.834
9.	10,372	7161.377
10.	4,082	6797.827

Deriving and Analyzing Regression Residuals

The next indispensable element of model diagnostics is the calculation of the **residuals**. Each residual quantifies the prediction error specific to an individual observation, representing the portion of the response variable that the established model failed to account for. Identifying patterns or extreme values within the residuals is often the first indication of model misspecification, the presence of influential outliers, or the violation of core regression assumptions.

Step 4: Generate the residuals (error terms).

We use the versatile `predict` command once more, but this time we must explicitly specify the `residuals` option to instruct Stata to calculate the error terms ($Y - hat{Y}$). We store these critical values in a new variable named **resid_price**:

predict resid_price, residuals

To achieve a comprehensive, record-by-record view of the model's performance, we list all three relevant variables together for the first few cases: the actual price, the predicted price, and the

calculated residual. This allows for immediate verification that the residual is indeed the precise difference between the observed outcome and the predicted outcome:

list price pred_price resid_price in 1/10

```
. predict resid_price, residuals

. list price pred_price resid_price in 1/10
```

	price	pred_p~e	resid_p~e
1.	4,099	5278.305	-1179.304
2.	4,749	7323.933	-2574.933
3.	3,799	5278.305	-1479.304
4.	4,816	6308.834	-1492.834
5.	7,827	8533.113	-706.1129
6.	5,788	6919.01	-1131.01
7.	4,453	6716.69	-2263.69
8.	5,189	6308.834	-1119.834
9.	10,372	7161.377	3210.623
10.	4,082	6797.827	-2715.827

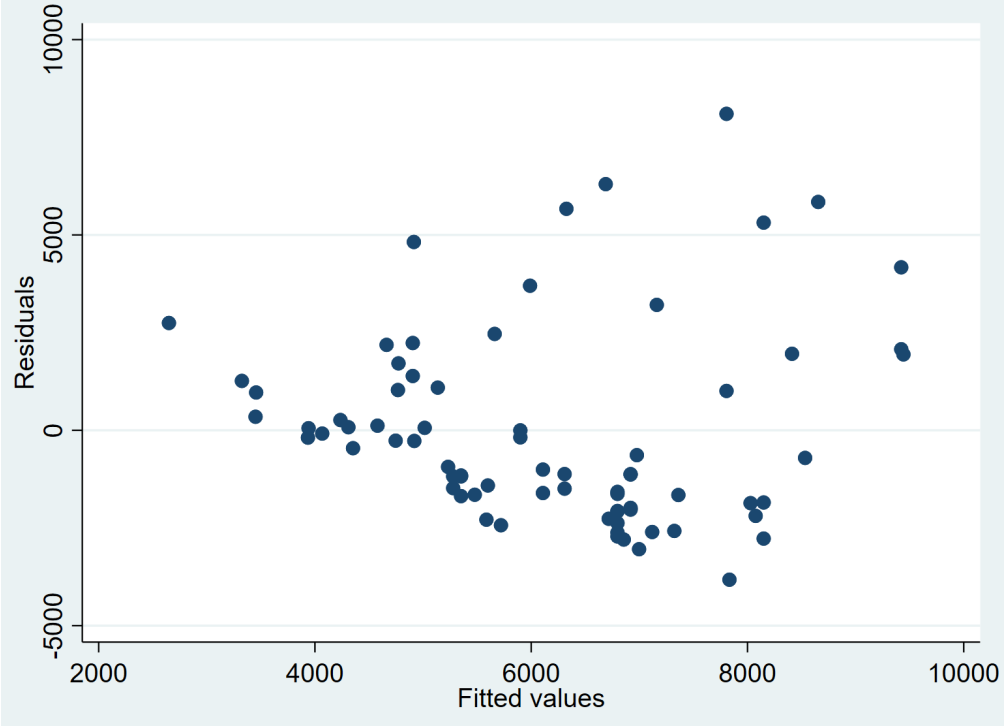
Essential Diagnostic Visualization: The Residual Plot

While examining numerical summaries of residuals is useful, graphical analysis provides the most powerful and intuitive method for checking the underlying assumptions of linear regression. The standard diagnostic visualization plots the residuals on the Y-axis against the predicted (fitted) values on the X-axis. For a model that perfectly meets the assumptions, the resulting scatterplot should display a completely random, uniform band of points centered around the horizontal zero line, with no discernible patterns or funnel shapes.

Step 5: Generate the predicted values versus residuals plot.

We use the `scatter` command in Stata to generate this crucial plot, specifying the residual variable first (Y-axis), followed by the predicted value variable (X-axis):

scatter resid_price pred_price



Upon visual examination of the resulting scatterplot, a clear and non-random pattern is evident: the spread, or variance, of the residuals systematically increases as the predicted prices (fitted values) become larger. This funnel shape indicates that the errors are not constant across all levels of the response variable, which is the classic visual signature of **heteroscedasticity**. This violates the core assumption of homoscedasticity (constant variance of errors), which is required for standard Ordinary Least Squares (OLS) estimation to be efficient.

Detecting heteroscedasticity through this visualization is paramount, as its presence leads to inaccurate standard errors, causing traditional hypothesis tests (like t-tests and F-tests) to be unreliable. While we identified the issue graphically, formal testing procedures, such as the White test, can confirm the violation statistically. The common solution is to utilize robust standard errors (or Huber-White standard errors) in the regression analysis, which adjusts the standard errors for the non-constant variance, thereby preserving the validity of statistical inference even when this fundamental assumption is compromised.