

Understanding Simple Linear Regression with Stata: A Comprehensive Tutorial

Authored by
Mohammed looti

November 8, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Understanding Simple Linear Regression with Stata: A Comprehensive Tutorial*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13721>

[Simple Linear Regression](#) (SLR) stands as a cornerstone technique in statistical analysis, fundamentally employed to quantify and model the straight-line relationship that exists between two distinct numerical variables. This relationship involves an [explanatory variable](#) (X), often referred to as the independent variable, and a corresponding [response variable](#) (Y), or dependent variable. Mastery of SLR is essential for anyone engaged in predictive modeling, assessing causal inference, or exploring correlations across diverse fields, ranging from social sciences to advanced engineering.

This tutorial provides a comprehensive, structured guide detailing the methodological and practical steps required to successfully execute and rigorously interpret a simple linear regression analysis. We will leverage the capabilities of the powerful statistical software, [Stata](#), guiding you through every phase of the process--from initial data preparation and critical visualization to formal model execution and the final, professional reporting of results. Our focus remains on generating clear, statistically sound insights from real-world data.

Case Study: Quantifying Automotive Fuel Efficiency

To provide a tangible demonstration of simple linear regression in practice, we will explore the relationship between the physical mass of an automobile and its corresponding fuel efficiency. Our primary objective is to quantify the precise connection between the car's weight (our explanatory variable) and its miles per gallon (MPG), which serves as our response variable. It is generally hypothesized that heavier vehicles require more energy to move, resulting in lower fuel economy; SLR provides the tools necessary to formally test and quantify the strength and direction of this inverse relationship.

For this hands-on analysis within Stata, we will utilize the widely known, internal sample dataset labeled *auto*. This dataset contains comprehensive specifications for 74 different vehicles, providing the necessary variables (*mpg* and *weight*) required for our analysis. The subsequent steps are organized logically to ensure a smooth progression from data access to the final statistical conclusion.

Step 1: Data Preparation and Initial Loading

The foundational step for any statistical modeling endeavor is ensuring the required data is correctly loaded and accessible within the analytical environment. While most real-world projects necessitate importing data from external formats (such as Excel spreadsheets or CSV files), Stata provides built-in sample datasets, which are ideal for pedagogical purposes and demonstration.

To load the *auto* dataset directly from the Stata Press online repository, execute the following command precisely in the Stata Command box. This action immediately makes the variables *mpg* and *weight* available for all subsequent analysis commands:

use <http://www.stata-press.com/data/r13/auto>

Following a successful load operation, it is considered best practice in data analysis to immediately generate an overview of the dataset's variables and their descriptive statistical properties. This preliminary step, often referred to as data vetting, confirms data integrity, helps identify potential issues like outliers or missing values, and verifies that the variables are of the appropriate type (e.g., continuous) for regression modeling.

Execute the subsequent command to produce a comprehensive summary of all variables contained within the newly loaded dataset:

summarize

```
. use http://www.stata-press.com/data/r13/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 that the dataset comprises 74 unique observations across 12 distinct variables. For the specific scope of our simple linear regression, our focus narrows entirely to the two critical continuous variables: *mpg* (miles per gallon) and *weight* (measured in pounds). The summary statistics, showing valid ranges and means, confirm that both variables are suitable for inclusion in a linear model.

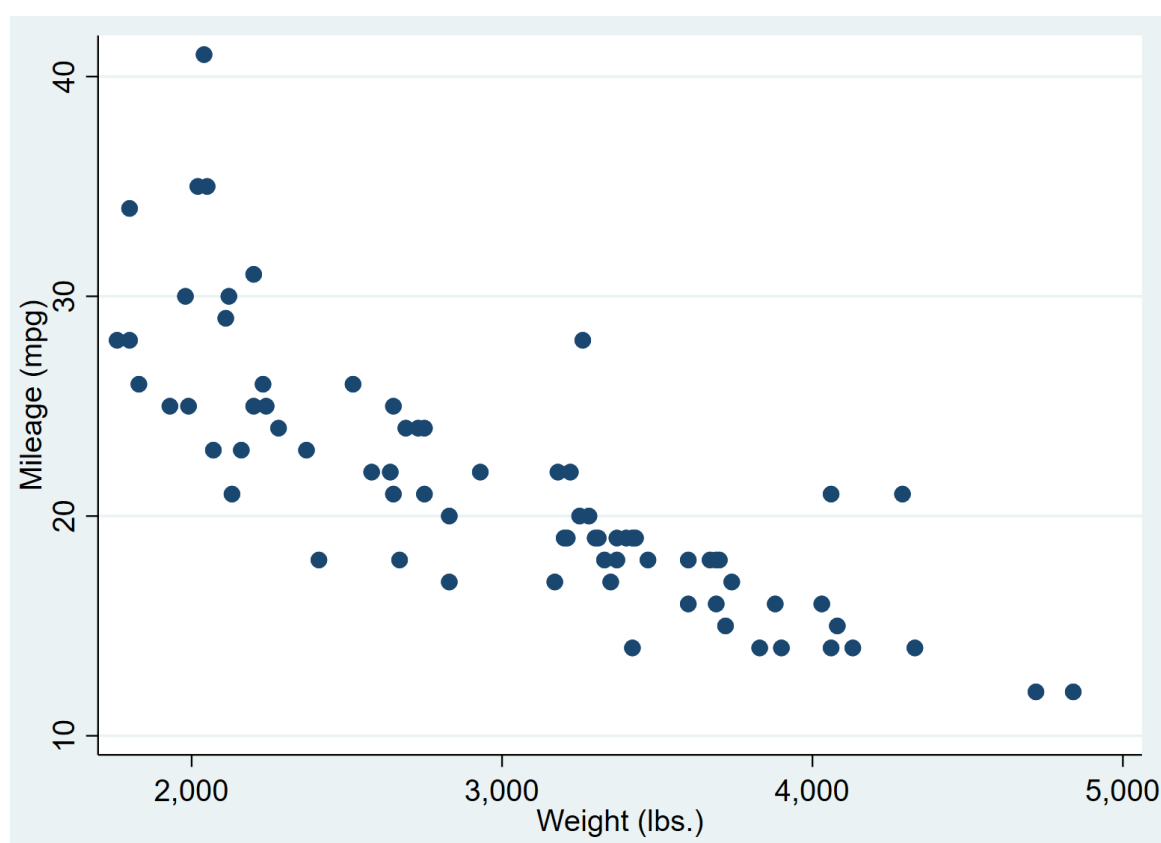
Step 2: Visualizing the Data Relationship

Prior to performing any formal statistical test, visualization serves as an indispensable step in regression analysis. A graphical representation allows the analyst to visually confirm assumptions, such as linearity, and to proactively identify structural patterns or anomalous data points (outliers) that might unduly influence the regression estimates. For evaluating the relationship between two continuous variables, the preferred and most informative visualization is the [scatterplot](#).

To generate a scatterplot that maps the relationship between car weight and fuel efficiency, input the following command into the Stata Command box. The conventional syntax requires listing the response variable (MPG) first, followed immediately by the explanatory variable (Weight):

```
scatter mpg weight
```

Executing this command yields the following graphical representation:



A thorough examination of the scatterplot reveals a distinct and strong negative linear trend. This visual evidence powerfully validates our initial hypothesis: as the weight of the automobile increases along the x-axis, the resulting miles per gallon consistently decreases along the y-axis. Crucially, the data points exhibit a tight clustering around what would be a downward-sloping line, providing strong preliminary confirmation that a simple linear model is the appropriate and most

effective tool for quantifying this observed inverse correlation.

Step 3: Executing the Formal Regression Model

With the linear nature of the relationship established through visualization, we are ready to move forward with the formal estimation of the simple linear regression model. The primary objective is to estimate the precise slope and intercept of the line of best fit--the line that minimizes the total sum of squared residuals (the vertical distances between the observed data points and the regression line itself).

To execute this analysis in Stata, we utilize the standard [Ordinary Least Squares](#) (OLS) estimation via the ``regress`` command. The command syntax requires specifying the response variable (MPG) followed by the explanatory variable (Weight).

regress mpg weight

. regress mpg weight

Source	SS	df	MS	Number of obs	=	74
Model	1591.9902	1	1591.9902	F(1, 72)	=	134.62
Residual	851.469256	72	11.8259619	Prob > F	=	0.0000
				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 resulting output generated by Stata is exhaustive, presenting crucial diagnostic statistics, ANOVA results, and, most importantly, the estimated [coefficients](#) for the model. The next critical step involves systematically decoding these numerical results to derive statistically sound conclusions regarding the nature and magnitude of the relationship between vehicle weight and fuel efficiency.

Step 4: Interpreting Key Model Statistics

The Stata regression summary provides several key metrics essential for defining the overall quality, significance, and predictive power of the estimated model. A thorough interpretation requires focused attention on the following core values:

R-squared (R-sq): The displayed R-squared value is 0.6515. This statistic, also known as the Coefficient of Determination, represents the proportion of the variance in the response variable (MPG) that is successfully explained by the explanatory variable (Weight). A value of 0.6515 means that 65.15% of the total variation observed in the miles per gallon across the 74 cars can be statistically accounted for or explained solely by differences in the car's weight. This value indicates a substantially strong fit for the linear model.

Coefficient (Coef) for Weight (Slope): The estimated coefficient for the variable *weight* is -0.0060087. This value represents the slope of the regression line and is the most important finding. It quantifies the average predicted change in the response variable (MPG) resulting from a one-unit increase in the explanatory variable (Weight). Practically, this means that for every one-pound increase in a car's weight, the average miles per gallon is predicted to decrease by approximately 0.006 units. This negative sign confirms the inverse relationship observed in the scatterplot.

Coefficient (Coef) for _cons (Intercept): The value 39.44028 represents the y-intercept. Mathematically, it is the predicted average value of MPG when the Weight variable is zero. While crucial for the mathematical definition of the line of best fit, its practical interpretation is nonsensical in this context, as a vehicle cannot possess zero weight.

P>|t| (P-value) for Weight: This value is 0.000. This is the p-value associated with the hypothesis test that the true population coefficient for weight is zero (i.e., that no relationship exists). Because our calculated p-value (0.000) is far smaller than the conventional significance threshold of 0.05, we possess overwhelming evidence to reject the null hypothesis. We conclude definitively that there is a **statistically significant** linear relationship between a car's weight and its miles per gallon.

Step 5: Forming and Utilizing the Predictive Equation

The ultimate practical outcome of a regression analysis is the formulation of a predictive equation that mathematically summarizes the quantified relationship between the variables. By incorporating the intercept and slope coefficients derived from the OLS estimation, we can formally construct the simple linear regression equation, which adheres to the structure: Predicted Y = Intercept + (Slope * X).

In the context of our fuel efficiency study, the final, estimated regression equation is precisely formulated as follows:

$$\text{predicted mpg} = 39.44028 - 0.0060087 * (\text{weight})$$

This powerful equation can now be used for practical prediction within the range of our observed data. If we wish to estimate the MPG for a hypothetical new car weighing exactly 4,000 pounds, we substitute this value into the derived equation:

$$\text{predicted mpg} = 39.44028 - 0.0060087 * (4000) = 15.405$$

Based on our model, a car weighing 4,000 pounds is expected to achieve an average fuel efficiency of approximately 15.405 miles per gallon. This ability to generate reliable, data-driven estimates represents the core utility of simple linear regression modeling.

Step 6: Formal Reporting of Findings

The final and crucial stage of any statistical project is the clear, accurate, and concise reporting of the results, typically formatted for an academic paper or a formal business report. A robust report must succinctly summarize the methodology, present the key statistical evidence (including significance and effect size), and clearly state the final predictive equation.

When reporting, it is mandatory to include the t-statistic and the associated p-value to provide statistical evidence for the claim of significance, alongside the R-squared value to convey the model's overall explanatory power. Below is a standardized and formal example of how the results of this simple linear regression analysis should be presented:

A simple linear regression analysis was executed to formally quantify the relationship between an automobile's weight (explanatory variable) and its miles per gallon (MPG, response variable), utilizing the specifications from a sample of 74 distinct vehicles.

The analysis demonstrated a statistically significant inverse linear relationship between car weight and MPG ($t = -11.60$, $p < 0.0001$). The model exhibited a strong fit, with the car's weight explaining 65.15% (**R-squared** = 0.6515) of the total observed variability in miles per gallon.

The estimated regression equation derived from the **Ordinary Least Squares** method was found to be:

$$\text{predicted mpg} = 39.44 - 0.006 * (\text{weight})$$

Interpretation of the slope coefficient (-0.006) indicates that, on average, every additional one-pound increase in the automobile's weight is predicted to result in a decrease of 0.006 miles per gallon in fuel efficiency.