

Understanding Multiple Linear Regression with Stata: A Practical Tutorial

Authored by
Mohammed looti

November 8, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Understanding Multiple Linear Regression with Stata: A Practical Tutorial*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13718>

Multiple linear regression is a powerful statistical technique employed to model the linear relationship between several predictor variables and a single outcome variable. It extends the concept of simple linear regression by allowing researchers to assess the influence of multiple factors simultaneously on the variable of interest, providing a more comprehensive understanding of complex phenomena.

This comprehensive tutorial is designed to guide you through the process of executing and interpreting a multiple linear regression analysis using the statistical software package, [Stata](#). We will cover everything from loading the dataset to interpreting the final coefficients and presenting the results professionally.

Understanding the Fundamentals of Multiple Linear Regression

Before diving into the software commands, it is crucial to establish a solid theoretical foundation. Multiple linear regression aims to fit a line (or hyperplane in higher dimensions) that best minimizes the sum of squared differences between the observed data points and the values predicted by the model. The resulting equation allows us to quantify how much the [explanatory variables](#) (also known as independent variables or predictors) contribute to the variation in the [response variable](#) (the dependent variable).

A key advantage of using MLR over running multiple simple regressions is that it controls for the effects of all included predictors simultaneously. This control is vital because, in real-world data, predictors are often correlated with one another. By including all relevant variables in one model, we can isolate the unique contribution of each predictor, providing a much clearer and less biased estimate of its true effect on the outcome.

For example, if we study the factors influencing house price, both size and location are important. A simple regression of price on size might overestimate the size effect if larger houses tend to be in better neighborhoods. MLR resolves this by estimating the effect of size while holding the effect of location constant, leading to more accurate inference.

Example Scenario: Modeling Automobile Price Drivers

For this tutorial, we will explore a practical example: determining if a car's fuel efficiency (miles per gallon, or *mpg*) and its physical mass (*weight*) influence its market *price*. To test this hypothesis, we will construct a multiple linear regression model using *mpg* and *weight* as our two explanatory variables and *price* as the response variable.

We will utilize the standard Stata built-in dataset, *auto*, which conveniently contains detailed specifications for 74 different automobiles. By following the subsequent steps, you will learn how to prepare the data, run the analysis, and interpret the crucial statistics generated by Stata.

The steps below outline the specific Stata commands required to execute this analysis using the *auto* dataset.

Executing the Analysis: Loading Data and Summary Statistics

Step 1: Load the Data.

To begin, you must load the dataset into Stata's active memory. This is achieved by typing the following command into the Command box, which retrieves the dataset directly from the Stata Press website:

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

Once the dataset is loaded, it is good practice to familiarize yourself with the structure and content of the variables you plan to use.

Step 2: Obtain a Summary of the Data.

To quickly understand the characteristics and data types of the variables involved, use the `summarize` command. Type the following into the Command box:

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 summary output indicates that the dataset contains 12 different variables. For our specific analysis, we are primarily concerned with three variables: *mpg*, *weight*, and the outcome variable, *price*. The summary statistics provide immediate insights into the scale and distribution of these key elements.

Based on the output, we can observe the following basic summary statistics for our three variables:

price: The average price (mean) is approximately \$6,165, ranging from a minimum of \$3,291 to a maximum of \$15,906.

mpg: The average fuel efficiency is 21.29 miles per gallon, with values spanning from 12 mpg to 41 mpg.

weight: The average car weight is 3,019 pounds, falling within the range of 1,760 pounds to 4,840 pounds.

These descriptive statistics confirm the data quality and prepare us for the inferential modeling phase.

Running the Multiple Regression Command

With the data loaded and summarized, we can now proceed to execute the multiple linear regression model. In Stata, the primary command for this procedure is `regress`. The syntax requires specifying the response variable first, followed by all explanatory variables.

Type the following command into the Command box to perform the multiple linear regression using *mpg* and *weight* as the predictors of *price*:

regress price mpg weight

Upon execution, Stata generates a comprehensive output table detailing the fit of the overall model and the estimates for the individual parameters. This output contains the necessary statistical evidence to test our hypotheses regarding the relationship between car characteristics and price.

```
. regress price mpg weight
```

Source	SS	df	MS	Number of obs	=	74
Model	186321280	2	93160639.9	F(2, 71)	=	14.74
Residual	448744116	71	6320339.67	Prob > F	=	0.0000
				R-squared	=	0.2934
				Adj R-squared	=	0.2735
Total	635065396	73	8699525.97	Root MSE	=	2514

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mpg	-49.51222	86.15604	-0.57	0.567	-221.3025	122.278
weight	1.746559	.6413538	2.72	0.008	.467736	3.025382
_cons	1946.069	3597.05	0.54	0.590	-5226.245	9118.382

The output is generally divided into three main sections: the ANOVA table (top left), the overall model fit statistics (top right), and the coefficient table (bottom). We must systematically interpret each section to draw valid conclusions from our analysis.

Interpreting Overall Model Fit and Global Significance

The overall model fit statistics provide an initial assessment of how well our chosen predictors collectively explain the variation in the response variable. Two key values to examine here are the F-statistic and the R-squared value.

Prob > F: The value 0.000 represents the P-value for the overall F-test (found in the ANOVA section). The F-test assesses the null hypothesis that all regression coefficients (excluding the intercept) are simultaneously equal to zero. Since this P-value (0.000) is significantly less than the conventional alpha level of 0.05, we reject the null hypothesis. This indicates that the combined set of explanatory variables (*mpg* and *weight*) has a **statistically significant** relationship with the response variable *price*. In simple terms, the model as a whole is useful for predicting price.

R-squared: The R-squared value is 0.2934. This statistic, often called the Coefficient of Determination, measures the proportion of the variance in the response variable that is predictable from the explanatory variables. In this specific example, 29.34% of the total variation observed in car prices can be explained by the linear combination of *mpg* and *weight*. While this leaves over 70% of the variation unexplained (suggesting other important factors like make, model, or age are missing), it still represents a measurable level of explanatory power.

It is important to note the difference between R-squared and Adjusted R-squared. While R-squared will always increase when more predictors are added (even non-significant ones), Adjusted R-squared penalizes the model for complexity, providing a more reliable measure of model

goodness-of-fit, especially when comparing models with different numbers of predictors.

Analyzing Individual Predictor Coefficients

The coefficient table provides the estimated impact of each individual predictor on the response variable, along with statistical tests to determine if that impact is significantly different from zero.

Coef (mpg): -49.512. This is the unstandardized regression [Coefficient](#) for miles per gallon. It indicates the average change in price associated with a one-unit increase in *mpg*, assuming that the other predictor, *weight*, is held constant. Specifically, each one-unit increase in *mpg* is associated with an average decrease of approximately \$49.51 in price. This counterintuitive negative relationship suggests that, once car weight is accounted for, better fuel economy might be associated with lower prices, perhaps because lighter, more efficient cars tend to be smaller or less luxurious.

To illustrate, consider two cars, A and B, both weighing exactly 2,000 pounds. If Car A achieves 20 mpg and Car B achieves 19 mpg, we would predict Car A (the more efficient one) to cost \$49.51 less than Car B, based on this model.

P>|t| (mpg): 0.567. This P-value is associated with the t-statistic for the *mpg* coefficient. Since 0.567 is substantially greater than the 0.05 threshold, we lack sufficient evidence to conclude that the coefficient for *mpg* is statistically different from zero. Therefore, while the coefficient suggests a negative relationship, we cannot confidently state that *mpg* has a statistically significant independent relationship with *price* when *weight* is included in the model.

Coef (weight): 1.746. This positive coefficient for *weight* indicates that for every one-unit increase in weight (one pound), the car's average price increases by approximately \$1.75, assuming *mpg* is held constant. This aligns with common knowledge, as heavier cars often correlate with larger size, more materials, and potentially greater safety or luxury features.

P>|t| (weight): 0.008. This P-value is for the *weight* coefficient. Since 0.008 is less than 0.05, we have strong statistical evidence to reject the null hypothesis. We can confidently conclude that *weight* has a **statistically significant** positive relationship with *price*, even after controlling for fuel efficiency.

Coef (_cons): 1946.069. This value represents the intercept (or constant) of the regression equation. Mathematically, it is the predicted price of a car when both *mpg* and *weight* are zero. While useful for constructing the algebraic regression equation ($\text{Price} = 1946.07 - 49.51 \cdot \text{mpg} + 1.75 \cdot \text{weight}$), interpreting the intercept in practical terms is often nonsensical, as a car cannot have zero weight or zero fuel efficiency. Its primary role is to properly anchor the regression line within the coordinate system defined by the data.

Step 4: Reporting and Contextualizing the Results

The final and most crucial step in any statistical analysis is communicating the findings clearly and accurately. The goal is to summarize the model's performance and articulate the specific, significant effects found.

Here is an example of how one might report the findings of this multiple linear regression analysis, ensuring all key statistics (coefficients, t-statistics, and P-values) are included:

Multiple linear regression was performed to quantify the relationship between the weight and fuel efficiency (mpg) of a car and its selling price. A sample of 74 cars was included in the analysis.

The overall model was statistically significant ($F(2, 71) = 14.65, p < .001$), explaining 29.34% of the variance in car price ($R\text{-squared} = .2934$). Results showed that car weight had a statistically significant positive relationship with price ($t = 2.72, p = .008$). Specifically, holding fuel efficiency constant, every additional pound of weight was associated with an average increase of \$1.75 in price. Conversely, fuel efficiency (mpg) did not demonstrate a statistically significant independent relationship with price ($t = -.57, p = 0.567$) when weight was controlled for.

This structured reporting format ensures that both the global fit of the model and the unique contributions of each predictor are clearly conveyed to the reader, fulfilling the objectives of the multiple linear regression analysis.