

# How to Create and Interpret a ROC Curve in Stata

Authored by  
**stats writer**

March 9, 2026

## RECOMMENDED CITATION

stats writer (2026). *How to Create and Interpret a ROC Curve in Stata*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=134761>

The process of creating and interpreting a ROC (receiver operating characteristic) curve in Stata involves several steps. First, the data must be organized into two groups, typically a "positive" group and a "negative" group. These groups can represent different outcomes, such as presence or absence of a disease, or success or failure of a treatment.

Next, a logistic regression model is used to calculate the probability of belonging to the positive group based on a set of predictor variables. This probability is then used to rank the observations from highest to lowest. A ROC curve is then created by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at different probability cutoffs. The area under the curve (AUC) is a measure of the model's overall accuracy, with a value of 0.5 indicating no predictive power and a value of 1 indicating perfect predictive power.

Interpreting the ROC curve involves examining the shape and position of the curve. A curve that is closer to the top left corner indicates a higher accuracy of the model, while a curve closer to the diagonal line indicates a lower accuracy. The optimal cutoff point can be determined by selecting the point on the curve closest to the top left corner, which maximizes both sensitivity and specificity.

Overall, the process of creating and interpreting a ROC curve in Stata allows for the evaluation of a model's predictive power and the determination of an optimal cutoff point for classification. It is a useful tool for assessing the performance of a logistic regression model.

## Create and Interpret a ROC Curve in Stata

**is a statistical method that we use to fit a regression model when the response variable is binary. To assess how well a logistic regression model fits a dataset, we can look at the following two metrics:**

**Sensitivity: the probability that the model predicts a positive outcome for an observation when indeed the outcome is positive. Specificity: the probability that the model predicts a negative outcome for an observation**

when indeed the outcome is negative.

One easy way to visualize these two metrics is by creating a ROC curve, which is a plot that displays the sensitivity and specificity of a logistic regression model.

This tutorial explains how to create and interpret a ROC curve in Stata.

**Example: ROC Curve in Stata**

For this example we will use a dataset called *lbw*, which contains the following variables for 189 mothers:

**low** - whether or not the baby had a low birthweight. 1 = yes, 0 = no.  
**age** - age of the mother.  
**smoke** - whether or not the mother smoked during pregnancy. 1 = yes, 0 = no.

We will fit a logistic regression model to the data using age and smoking as explanatory variables and low birthweight as the response variable. Then we will create a ROC curve to analyze how well the model fits the data.

## Step 1: Load and view the data.

Load the data using the following command:

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

Gain a quick understanding of the dataset using the following command:

**summarize**

```
. use http://www.stata-press.com/data/r13/lbw
(Hosmer & Lemeshow data)
```

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	189	121.0794	63.30363	4	226
low	189	.3121693	.4646093	0	1
age	189	23.2381	5.298678	14	45
lwt	189	129.8201	30.57515	80	250
race	189	1.846561	.9183422	1	3
smoke	189	.3915344	.4893898	0	1
ptl	189	.1957672	.4933419	0	3
ht	189	.0634921	.2444936	0	1
ui	189	.1481481	.3561903	0	1
ftv	189	.7936508	1.059286	0	6
bwt	189	2944.286	729.016	709	4990

There are 11 different variables in the dataset, but the only three that we care about are low, age, and smoke.

## Step 2: Fit the logistic regression model.

Use the following command to fit the logistic regression model:

## logit low age smoke

```
. logit low age smoke
```

```
Iteration 0: log likelihood = -117.336
Iteration 1: log likelihood = -113.66733
Iteration 2: log likelihood = -113.63815
Iteration 3: log likelihood = -113.63815
```

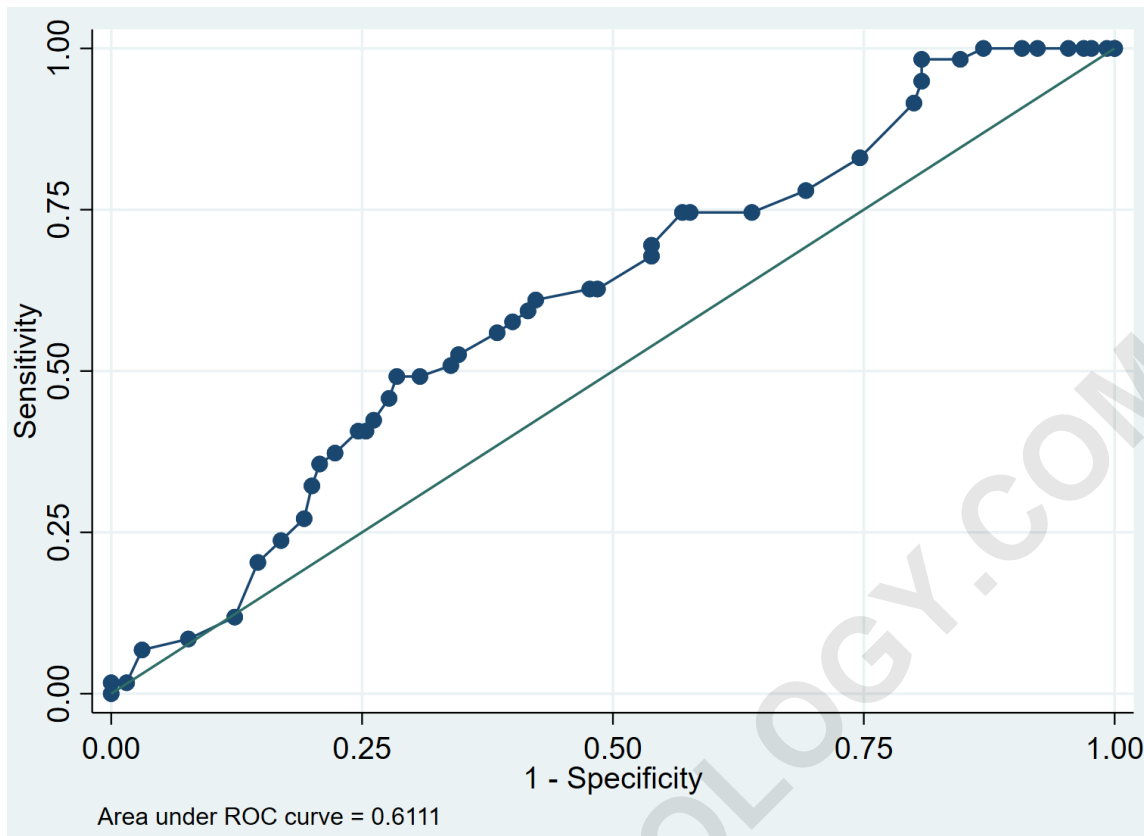
```
Logistic regression              Number of obs   =      189
                                LR chi2(2)      =       7.40
                                Prob > chi2         =     0.0248
Log likelihood = -113.63815      Pseudo R2      =     0.0315
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	-.0497792	.031972	-1.56	0.119	-.1124431 .0128846
smoke	.6918486	.3218061	2.15	0.032	.0611202 1.322577
_cons	.0609051	.7573199	0.08	0.936	-1.423415 1.545225

## Step 3: Create the ROC curve.

We can create the ROC curve for the model using the following command:

## lroc



#### Step 4: Interpret the ROC curve.

When we fit a logistic regression model, it can be used to calculate the probability that a given observation has a positive outcome, based on the values of the predictor variables.

To determine if an observation should be classified as positive, we can choose a cut-point such that observations with a fitted probability above the cut-point are classified as positive and any observations with a fitted probability below the cut-point are

**classified as negative.**

**For example, suppose we choose the cut-point to be 0.5. This means that any observation with a fitted probability greater than 0.5 will be predicted to have a positive outcome, while any observation with a fitted probability less than or equal to 0.5 will be predicted to have a negative outcome.**

**The ROC curve shows us the values of sensitivity vs. 1-specificity as the value of the cut-off point moves from 0 to 1. A model with high sensitivity and high specificity will have a ROC curve that hugs the top left corner of the plot. A model with low sensitivity and low specificity will have a curve that is close to the 45-degree diagonal line.**

**The AUC (area under curve) gives us an idea of how well the model is able to distinguish between positive and negative outcomes. The AUC can range from 0 to 1. The higher the AUC, the better the model is at correctly classifying outcomes. In our example, we can see that the AUC is 0.6111.**

**We can use AUC to compare the performance of two or**

**more models. The model with the higher AUC is the one that performs best.**

ARABPSYCHOLOGY.COM