A Practical Guide to ROC Curve Analysis and Interpretation in Stata for Logistic Regression

Authored by

Mohammed looti

November 8, 2025

Logistic regression is a fundamental statistical technique employed when the dependent variable, or response, is a categorical variable restricted to exactly two possible outcomes. This scenario is widely known as binary classification. The core objective of this modeling approach is to estimate the probability of a specific event occurring, given a set of predictor variables.

To properly assess the efficacy and predictive capability of any fitted logistic regression model, particularly in high-stakes classification tasks, we must rely on specialized performance metrics. The two most critical metrics for evaluating how well a model discriminates between classes are:

Sensitivity (True Positive Rate): This metric measures the proportion of actual positive outcomes that the model correctly identified. High sensitivity is paramount when the analytical priority is minimizing **false negatives**--cases where a positive event is incorrectly missed.
Specificity (True Negative Rate): This metric measures the proportion of actual negative outcomes that were correctly identified. High specificity is essential when the goal is to minimize **false positives**--cases where a negative event is incorrectly classified as positive.

A powerful graphical tool for visualizing the inherent trade-off between these two metrics across all possible classification cut-points is the **ROC curve** (Receiver Operating Characteristic curve). This visualization provides a comprehensive, threshold-independent summary of the model's overall discriminatory performance.

This comprehensive tutorial details the necessary steps for fitting a logistic regression model, generating the associated ROC curve, and accurately interpreting the resulting statistics within the popular statistical software package, **Stata**.

Understanding Binary Outcomes and the Need for Logistic Regression

When researchers or analysts work with data where the outcome variable is dichotomous--such as disease presence/absence, customer churn/retention, or success/failure--standard linear regression methods are inappropriate. This is because linear models assume a continuous, normally distributed error term, an assumption fundamentally violated by a binary dependent variable.

Logistic regression overcomes this limitation by utilizing the logistic function (or sigmoid function) to transform the linear combination of predictor variables into a value ranging from zero to one. This resulting value is interpreted as the probability that the observation belongs to the positive class. The relationship modeled is non-linear, making it perfectly suited for predicting probabilities.

Although the model generates a continuous probability for each observation, in practical application, we must convert these probabilities into a discrete classification decision (Positive or Negative). This conversion requires establishing a **threshold** or cut-point (commonly 0.5). The

selection of this specific cut-point is crucial, as it fundamentally dictates the balance between the model's **sensitivity** and **specificity**. Understanding how performance shifts across all potential thresholds is exactly why the ROC curve is indispensable.

Visualizing Model Performance: The Mechanics of the ROC Curve

The ROC curve serves as a diagnostic plot used to evaluate the performance of a classifier system. It is constructed by plotting the True Positive Rate (Sensitivity) on the Y-axis against the False Positive Rate (1 - Specificity) on the X-axis. Each singular point along the generated curve represents a different classification threshold that could be used to convert the predicted probabilities into binary outcomes.

The plot visually captures the performance characteristics of the classifier as the discrimination threshold is systematically varied from 0 to 1. An ideal, theoretically perfect model would achieve 100% sensitivity (no false negatives) and 100% specificity (no false positives). Graphically, this perfect performance would manifest as a curve that passes directly through the top-left corner of the plot, where the True Positive Rate is 1 and the False Positive Rate is 0.

Conversely, a classifier that performs no better than assigning classifications randomly would yield a diagonal line, often referred to as the line of no-discrimination (running from (0,0) to (1,1)). Any useful and valid classification model must produce a curve that bows significantly above this diagonal benchmark, indicating performance superior to chance. The greater the curvature and the closer the curve tracks toward the top-left corner, the stronger the model's overall discriminatory power is considered.

Preparing Data for ROC Analysis in Stata

To practically demonstrate the process of fitting a logistic regression and generating the ROC curve, we will utilize the publicly available *lbw* dataset. This dataset is a standard example in statistical learning, containing records for 189 mothers and is used to predict the risk of low birthweight based on various maternal factors.

Our specific analytical objective is to fit a logistic regression model using two key predictor variables--maternal age and smoking status--to predict the dichotomous response variable, **low birthweight**. This setup allows us to assess the joint predictive utility of these factors.

The essential variables within the *lbw* dataset for this analysis are defined as follows:

low - This is the **binary response variable**. A value of 1 indicates the baby had a low birthweight, and 0 indicates a normal birthweight.
age - A continuous variable representing the age of the mother in years at the time of delivery.

smoke - A binary indicator variable showing whether the mother smoked during pregnancy (1 = yes, 0 = no).

We initiate the process by loading this dataset directly into our Stata session using a direct URL command. This method ensures reproducibility and easy access to the required data.

Executing the Logistic Model and Generating the ROC Plot in Stata

The first operational step in Stata is always to retrieve and load the necessary data file into the active session. We use the following command to fetch the dataset directly from the Stata Press repository:

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

Once the data is loaded, it is considered best practice to execute a summary command. This provides a quick check of the variable types, confirms the sample size (N=189), and offers basic descriptive statistics before proceeding with the complex modeling steps.

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

Though the dataset contains numerous variables, our focus is restricted to predicting *low* birthweight based on *age* and *smoke* status. With the data confirmed, we proceed to fit the logistic regression model using the standard `logit` command. We specify the outcome variable first, followed by the predictor variables:

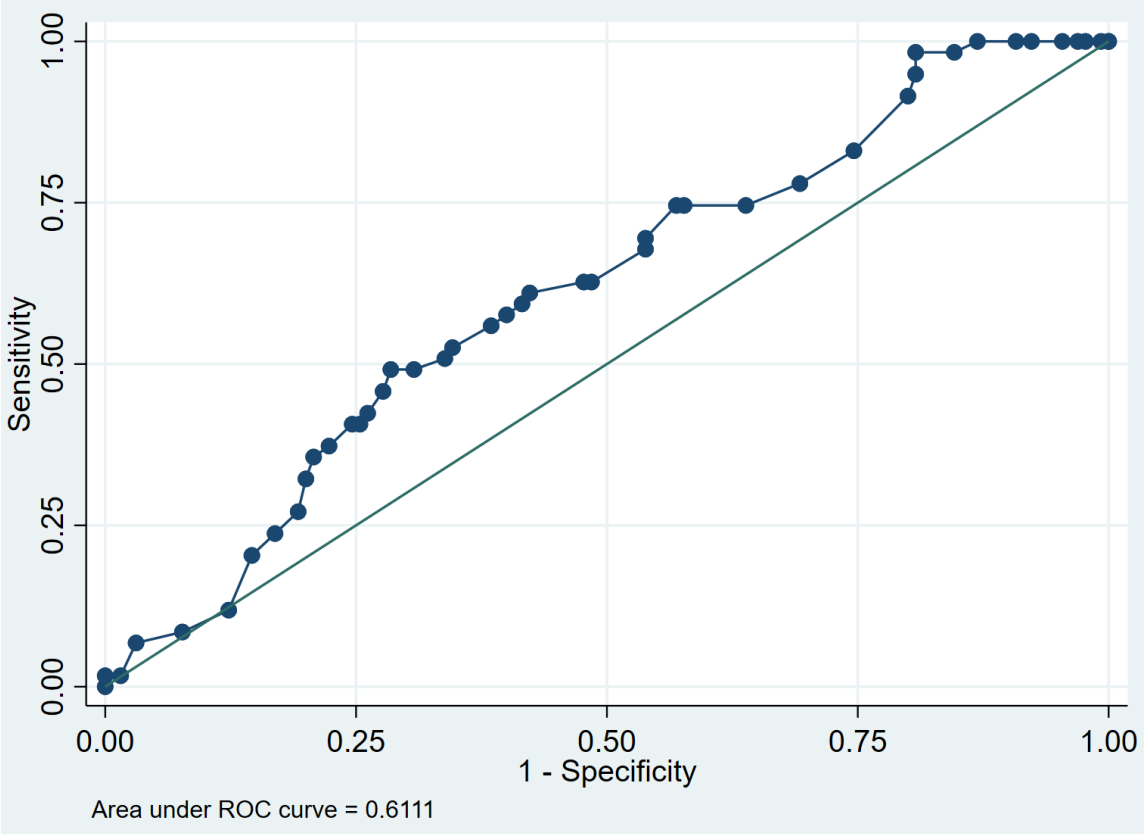
logit low age smoke

```
. 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
------------------------------------------------------------------------------
```

The resulting output confirms that the model has been successfully fitted. Crucially, Stata has internally calculated the predicted probabilities for every observation based on the estimated coefficients. These probabilities form the fundamental basis for generating the ROC curve. To visualize this performance profile, we use the dedicated ROC command, `lroc`, which must be executed immediately following the `logit` command, as it operates on the results of the most recently fitted binary outcome model.

We generate the ROC curve for the fitted model using the following concise command:

lroc



Interpreting the ROC Plot and Area Under the Curve (AUC)

The output generated by the `lroc` command provides two essential pieces of information: the visual representation of the curve itself and a crucial calculated metric known as the **Area Under the Curve (AUC)**.

The shape of the curve graphically communicates the model's overall discriminatory ability. As established, a strong model will generate a curve that "hugs" the upper-left corner of the plot. This shape signifies maximizing the Sensitivity (True Positive Rate, Y-axis) while simultaneously minimizing the False Positive Rate (X-axis). The steeper the initial ascent of the curve, the more effectively the model is separating the positive class observations from the negative class observations.

The concept of the **cut-point** is fundamental to interpreting the ROC curve. When performing classification, an analyst must select a specific cut-point (e.g., 0.5) for the predicted probability. Observations with a fitted probability above this point are classified as positive, and those below are negative. The strength of the ROC curve is that it visualizes how Sensitivity and Specificity react and change as this cut-point hypothetically moves across the entire range of probabilities from 0 to 1, providing a holistic view of performance.

However, the most definitive statistical measure derived from the plot is the **AUC (Area Under the Curve)** statistic. The AUC quantifies the overall performance of the classifier and is interpreted probabilistically: it represents the probability that the model will rank a randomly chosen positive observation higher (assign it a higher predicted probability) than a randomly chosen negative observation.

The AUC ranges strictly from 0 to 1. An AUC of 0.5 indicates that the model's performance is indistinguishable from random chance, possessing no discriminatory power whatsoever. Conversely, an AUC of 1.0 represents a perfect model capable of flawless classification across all thresholds. General guidelines suggest that an AUC between 0.7 and 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and 0.9 or higher is outstanding.

In our specific example, the Stata output calculates the AUC for the logistic regression model predicting low birthweight as **0.6111**. This value suggests that our model performs discernibly better than random chance (0.5), but its discriminatory power is relatively weak. This interpretation indicates that maternal age and smoking status alone are only moderately predictive of low birthweight in this particular dataset, suggesting that additional clinical or demographic predictors may be required to achieve a truly robust classification model.

Summary: Validating Model Performance

Creating and meticulously interpreting the ROC curve in Stata is a vital, non-negotiable step in validating the performance of any binary classification model. By utilizing the standard `logit` command followed by the diagnostic `lroc` command, practitioners can efficiently visualize the crucial trade-off between sensitivity and specificity and obtain the critical, threshold-independent AUC value.

A thorough understanding of these metrics ensures that the chosen model is robust, reliable, and suitable for deployment, especially in specialized fields like medicine, engineering, or finance where the relative cost associated with false positives versus false negatives must be carefully assessed and balanced. Analysts should always strive for a model whose ROC curve maximizes the area above the diagonal line, as this directly translates to a higher AUC and, consequently, superior overall predictive capability.