Learning Logistic Regression with SAS: A Step-by-Step Guide

Authored by

Mohammed looti

November 1, 2025

Understanding the Foundation of Logistic Regression

Logistic regression stands as a fundamental statistical method used extensively when the objective is to model the relationship between predictor variables and a response variable that is binary or dichotomous. Unlike traditional linear regression, which predicts a continuous outcome, logistic regression estimates the probability that an event will occur (e.g., 1 for success, 0 for failure). This powerful technique is indispensable across fields like medicine, finance, and social sciences for predictive modeling and classification tasks.

Since probabilities must fall between zero and one, logistic regression employs a link function--specifically the logit function--to transform the probability into a continuous scale. This transformation ensures that the model output is not constrained by the 0 to 1 boundary, allowing the linear combination of predictors to be used effectively. The coefficients of the model are determined using a computational process known as maximum likelihood estimation (MLE), which iteratively seeks the coefficient values that maximize the likelihood of observing the actual outcome data.

The core functional form of the logistic regression model results in the following equation, which models the linear relationship between the predictors and the natural log odds of the response variable being equal to 1:

log = β0 + β1X1 + β2X2 + ... + βpXp

In this formulation, the components are defined precisely to relate the predictors to the transformed probability:

Xj: Represents the jth predictor variable included in the model.
βj: Corresponds to the coefficient estimate for the jth predictor variable, quantifying the change in the log odds associated with a one-unit change in Xj.

The expression on the right-hand side predicts the **log odds** of the response variable taking on the designated "success" value (typically 1). The following comprehensive, step-by-step example demonstrates the practical application of fitting and interpreting a logistic regression model within the powerful statistical software environment of **SAS (Statistical Analysis System)**.

Step 1: Preparing and Structuring the Dataset in SAS

The initial stage of any statistical analysis involves meticulous data preparation. In SAS, this often means utilizing the `DATA` step to define and populate the variables that will be used in the model. For this demonstration, we will construct a hypothetical dataset tracking college acceptance criteria for 18 distinct students. This dataset includes a binary outcome variable and two continuous

predictor variables.

The three essential variables collected for each student are:

Acceptance into a certain university (coded as 1 for acceptance, 0 for denial).
GPA (recorded on a standard scale from 1.0 to 4.0).
ACT score (ranging from 1 to 36).

The following SAS code block executes the creation of the dataset named `my_data` and subsequently employs `PROC PRINT` to display the contents, ensuring the data structure is correct before proceeding to the modeling phase.

```
/*create dataset*/
data my_data;
input acceptance gpa act;
datalines;
1 3 30
0 1 21
0 2 26
0 1 24
1 3 29
1 3 34
0 3 31
1 2 29
0 1 21
1 2 21
0 1 15
1 3 32
1 4 31
1 4 29
0 1 24
1 4 29
1 3 21
1 4 34
;
run;

/*view dataset*/
proc print data=my_data;
```

Once the code is executed, SAS generates the initial output table, confirming that all 18 observations have been loaded correctly, associating the acceptance status with the corresponding GPA and ACT scores. This visual verification is a critical checkpoint before moving on to the complex computations involved in regression modeling.

Obs	acceptance	gpa	act
1	1	3	30
2	0	1	21
3	0	2	26
4	0	1	24
5	1	3	29
6	1	3	34
7	0	3	31
8	1	2	29
9	0	1	21
10	1	2	21
11	0	1	15
12	1	3	32
13	1	4	31
14	1	4	29
15	0	1	24
16	1	4	29
17	1	3	21
18	1	4	34

Step 2: Implementing the Logistic Regression Model using PROC LOGISTIC

With the dataset successfully prepared, the next step is to invoke the specialized procedure for fitting logistic models: **PROC LOGISTIC**. This procedure is the engine within SAS that performs the iterative maximum likelihood estimation required to determine the optimal coefficient estimates. We specify the dataset, the response variable, and the set of continuous predictor variables.

In this specific instance, we designate "acceptance" as the binary response variable, which the model attempts to predict, and "gpa" and "act" as the independent predictor variables. The syntax is concise and powerful, allowing SAS to manage the complex calculations internally.

A crucial command option in the `PROC LOGISTIC` statement is **DESCENDING**. By default, SAS attempts to model the probability of the lowest ordered response level (often 0). However, in logistic regression, we almost always wish to model the probability of the "success" event (coded

as 1). The inclusion of the **DESCENDING** option explicitly instructs SAS to model the probability that the response variable (`acceptance`) will take on the value of 1, ensuring the resulting coefficient interpretations align with the desired outcome prediction. Failure to include this option would result in the model predicting P(Y=0), potentially leading to inverse interpretations of the results.

```
/*fit logistic regression model*/
proc logistic data=my_data descending;
model acceptance = gpa act;
run;
```

Upon execution, SAS generates extensive output, including several tables that must be carefully analyzed to assess the overall model performance and the specific impact of the predictor variables. The initial output summarizes the model's configuration and provides the iterative history of the maximum likelihood estimation process.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	26.057	16.595
SC	26.947	19.266
-2 Log L	24.057	10.595

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13.4620	2	0.0012
Score	10.5311	2	0.0052
Wald	5.2807	2	0.0713

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.2839	4.2665	0.5924	0.4415
gpa	1	2.9665	1.6250	3.3324	0.0679
act	1	-0.1145	0.2369	0.2336	0.6289

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gpa	19.423	0.804	469.398
act	0.892	0.561	1.419

Interpreting the SAS Output: Model Fit Statistics

The SAS output provides several key metrics to evaluate how well the fitted model explains the variation in the data compared to a null model (a model with no predictors). These statistics are crucial for determining the overall utility and statistical soundness of the logistic regression.

One of the first critical measures provided is the AIC (Akaike Information Criterion). For our model, the AIC value is reported as **16.595**. The AIC serves as an estimate of the relative quality of statistical models for a given set of data; it is essentially a measure that balances model fit and model complexity. The fundamental principle is that a lower AIC value indicates a superior model fit. It is important to note that AIC does not provide an absolute measure of quality; rather, it is used primarily for comparative analysis. When comparing several competing models fitted to the same dataset, the model exhibiting the lowest AIC is conventionally considered the most parsimonious

and effective choice.

Following the AIC, the table titled **Testing Global Null Hypothesis: BETA=0** is essential for assessing the model's collective significance. This test evaluates the null hypothesis that all regression coefficients (excluding the intercept) are simultaneously equal to zero, implying that the predictor variables, as a group, have no effect on the response variable. This is analogous to the overall F-test in linear regression.

From this table, we observe the Likelihood Ratio Chi-square value, which is calculated as **13.4620**, corresponding to a degrees of freedom equal to the number of predictors (2 in this case). Critically, the associated p-value is **0.0012**. Since this p-value is substantially less than the conventional significance level of 0.05, we reject the null hypothesis. This powerful finding confirms that the logistic regression model, incorporating both GPA and ACT score, is statistically significant in its ability to predict college acceptance.

Interpreting the SAS Output: Coefficient Estimates and Significance

The most detailed and interpretable section of the output is presented in the table titled **Analysis of Maximum Likelihood Estimates**. This table provides the numerical value for each estimated coefficient, along with standard errors, Wald Chi-square statistics, and corresponding p-values for individual predictors. These estimates are central to understanding the impact and directionality of each variable on the log odds of the outcome.

The estimated coefficients quantify the average change in the log odds of the response variable (acceptance = 1) for a one-unit increase in the respective predictor variable, holding all other variables constant. Analyzing the coefficients for GPA and ACT score provides specific insights into their influence on university acceptance:

The coefficient for **GPA** is **2.9665**. This positive value indicates that a one-unit increase in GPA is associated with an average increase of **2.9665** in the log odds of being accepted into the university.
The coefficient for **ACT score** is **-0.1145**. This negative value indicates that a one-unit increase in ACT score is associated with an average *decrease* of **0.1145** in the log odds of acceptance, suggesting a counter-intuitive or non-significant relationship within this specific dataset.

While the coefficient signs suggest the direction of the effect, the statistical significance of each individual predictor must be assessed using its corresponding p-value provided in the table. The p-value for each coefficient tests the null hypothesis that that specific coefficient is zero, meaning the predictor has no significant effect on the log odds of the outcome when controlling for other variables in the model.

The individual p-values lead to critical conclusions regarding the predictive power of each variable:

The P-value associated with **GPA** is **0.0679**. While this value is slightly greater than the typical 0.05 threshold, it is very close, suggesting GPA is a marginally or near-statistically significant predictor of university acceptance. Further analysis or a slightly larger sample size might confirm its significance.
The P-value associated with **ACT score** is **0.6289**. Since this value is considerably higher than 0.05, we fail to reject the null hypothesis for the ACT score coefficient. This strongly suggests that, within the context of this model and in conjunction with GPA, the ACT score is not a statistically significant predictor of university acceptance.

In summary, the model suggests that GPA is the primary driver of acceptance probability, exerting a strong positive influence on the log odds, whereas the ACT score does not provide statistically significant additional predictive power when GPA is already included in the model. This detailed analysis allows practitioners to refine their models by potentially removing non-significant predictors or exploring interaction terms.

Additional Resources for SAS Regression Analysis

Exploring various regression methodologies enhances one's ability to model complex data structures effectively. The following resources offer deeper dives into how to implement other specialized regression models within the SAS environment, building upon the foundational knowledge gained from logistic regression: